JP9-2002-0132-US4

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 10-254883
(43)Date of publication of application : 25.09.1998

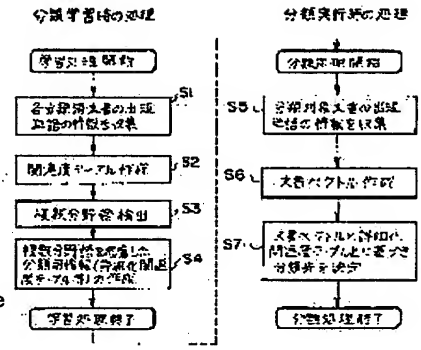
(51)Int.Cl. G06F 17/30

(21)Application number : 09-054359
(22)Date of filing : 10.03.1997

(71)Applicant : MITSUBISHI ELECTRIC CORP
(72)Inventor : FUJII YOICHI
SUZUKI KATSUSHI
IMAMURA MAKOTO
TAKAYAMA YASUHIRO

(54) AUTOMATIC DOCUMENT SORTING METHOD

(57)Abstract:
PROBLEM TO BE SOLVED: To provide an automatic document sorting method capable of precisely sorting even for detailed sorting.
SOLUTION: For learning, a word dividing/frequency extracting part collects information on appearing words from each sorted document (S1). A relation degree arithmetic part finds a relation degree between each word and each sort based on this information (S2). A plural-fields word processing part detects plural-fields word related with plural fields from this relation degree table (S3), divides each plural fields word by each strongly related field to regard as separated words and prepares information for sorting in a detailed relation degree table, etc., (S4). In sorting a document, a word dividing/frequency extracting processing part first collects information on the frequency, etc., of appearing words in the document (S5). A sorting destination deciding part 10 prepares a document vector expressing the tendency of appearing words in the sorting object document based on this information (S6) and decides the sorting destination of the document, based on this vector and the detailed relation degree table (S7).



LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

(19)日本国特許庁（ＪＰ）　(12) **公開特許公報（Ａ）**　(11)特許出願公開番号

特開平10－254883

(43)公開日　平成10年(1998)９月25日

(51)Int.Cl.[6]	識別記号	Ｆ Ｉ	
Ｇ０６Ｆ　17/30		Ｇ０６Ｆ　15/401	３１０Ｄ

審査請求　未請求　請求項の数４　ＯＬ　（全 27 頁）

(21)出願番号　特願平9－54359

(22)出願日　平成９年(1997)３月10日

(71)出願人　000006013
三菱電機株式会社
東京都千代田区丸の内二丁目２番３号
(72)発明者　藤井　洋一
東京都千代田区丸の内二丁目２番３号　三菱電機株式会社内
(72)発明者　鈴木　克志
東京都千代田区丸の内二丁目２番３号　三菱電機株式会社内
(72)発明者　今村　誠
東京都千代田区丸の内二丁目２番３号　三菱電機株式会社内
(74)代理人　弁理士　吉田　研二　（外２名）

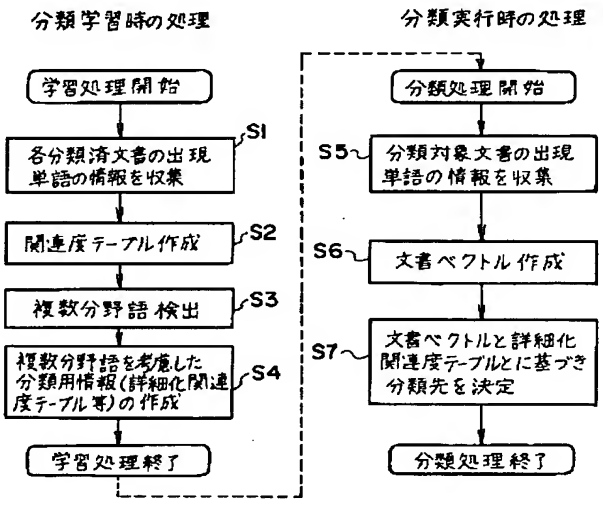
最終頁に続く

(54)【発明の名称】　文書自動分類方法

(57)【要約】

【課題】　細かい分類に対しても精度のよい分類を行うことができる文書自動分類方法を提供する。

【解決手段】　学習時には、単語分割／頻度抽出部は各分類済み文書から出現単語の情報を収集する（Ｓ１）。この情報に基づき関連度演算部が各単語と各分類との関連度を求め、関連度テーブルを作成する（Ｓ２）。複数分野語処理部は、この関連度テーブルから複数の分野に対して関連の強い複数分野語を検出し（Ｓ３）、各複数分野語を関連の強い各分野ごとに分割して別々の単語とみなして、詳細化関連度テーブルなどの分類用情報を作成する（Ｓ４）。文書を分類する際には、まず単語分割／頻度抽出処理部３が、当該文書の出現単語の頻度等の情報を収集する（Ｓ５）。分類先決定部１０は、この情報に基づき当該分類対象文書の出現単語の傾向を表す文書ベクトルを作成し（Ｓ６）、このベクトルと詳細化関連度テーブルとに基づき当該文書の分類先を決定する（Ｓ７）。



1

【特許請求の範囲】

【請求項１】　分類済みの各文書に出現する各単語の頻度集計結果に基づき各単語と各分野との関連度を登録した関連度テーブルを作成し、この関連度テーブルから、閾値より高い関連度を有する強関連分野が複数存在する複数分野単語を求め、前記関連度テーブルにおける複数分野語についての欄を、当該複数分野語とこれに対応する強関連分野との組合せごとに複数の欄に分割して詳細化関連度テーブルを作成する分類学習ステップと、

分類対象の文書に出現する単語の頻度を集計し、この結果得られた頻度情報を前記複数分野語の情報によって詳細化し、この詳細化された頻度情報と前記詳細化関連度テーブルとに基づき当該文書の分類先の分野を決定する分類実行ステップと、を含むことを特徴とする文書自動分類方法。

【請求項２】　単語の前記強関連分野の判定基準となる前記閾値は、前記関連度テーブルにおける当該単語の各分野に対する関連度の中の最大値に基づき定められることを特徴とする請求項１記載の文書自動分類方法。

【請求項３】　前記分類学習ステップでは、各複数分野語について、当該複数分野語の各強関連分野ごとに、その強関連分野に属する分類済み文書において当該複数分野語と共起した単語の傾向を表す共起ベクトルを生成し、

前記分類実行ステップでは、各複数分野語ごとに、この分類対象文書において当該複数分野語と共起した単語の傾向を示す文書共起ベクトルを生成し、この文書共起ベクトルと前記各共起ベクトルとの類似性に基づき、分類対象文書から得られた前記頻度情報を詳細化することを特徴とする請求項１又は２記載の文書自動分類方法。

【請求項４】　各単語の概念的な階層関係を記述したシソーラスを利用して、前記各共起単語の上位概念の情報を反映した共起ベクトル及び文書共起ベクトルを生成することを特徴とする請求項３記載の文書自動分類方法。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】この発明は、文書を自動分類する文書自動分類方法に関し、特に文書に出現する各単語の頻度の情報に基づき文書の分類先を決定する文書自動分類方法に関する。

【０００２】

【従来の技術】文書自動分類の方式のなかの有力なものの一つに、分類先の分野が既知の文書から出現単語の頻度統計をとって各分野に固有のキーワードや各分野の単語出現傾向などを学習し、これらを分類基準として用いて文書を分類するという方式がある。

【０００３】このような方式の文書自動分類装置の一つに、例えば特開平６－３４８７５５号公報に示される装置がある。この装置では、分類済みの文書群から各分野に固有の単語（キーワード）を抽出し、分類対象文書の

2

中のこれらキーワードの頻度に基づき当該文書の分類先を決定している。

【０００４】図３２は、この従来装置の構成図である。以下、分野のキーワードを学習する学習処理と、分類対象文書を分類する分類処理とに分けて、この従来装置の構成及び機能について説明する。

【０００５】まず、学習処理では、分類済み文書データ１０１から全ての文書が取り出され、文書データ単語分割部１０３で単語分割がされ、この分割結果の情報が分類済み文書単語分割テーブル１０４に格納される。分類用辞書作成部１０６は、分類済み文書単語分割テーブル１０４に格納された単語分割テーブルの情報に基づき、特定の分野のみに現れる単語を当該分類のキーワードとして抽出し、これらキーワードの（見出し文字列、品詞、分野名）の組を分類用辞書１０７に格納する。このようにして分類用辞書１０７が完成すると、学習処理は終了する。図３３は、この学習処理の結果得られた分類用辞書１０７の内容例である。この例は、例えば単語「自然語」は、分類名「言語処理」のキーワードであり、学習した分類済み文書群の「題名」に０回、「要旨」に１回、「目的」に１回、その他合計して１７回出現したことを示している。

【０００６】次に分類処理では、分類対象文書データ１０２から取り出された分類対象文書に対して、学習時と同様に文書データ単語分割部１０３で単語分割が行われ、この結果得られた単語分割データが分類対象文書単語分割テーブル１０５に格納される。文書分類部１１０は、分類対象文書単語分割テーブル１０５に格納された分類対象文書の単語分割データと、分類用辞書１０７の（見出し文字列、品詞、分野名）との間で、同じ単語を含むものの一致回数を各「分野名」毎に集計し、一致回数の最も多い「分野名」を当該分類対象文書の分類先の最優先候補として文書分類結果１１１に格納する。

【０００７】この文書分類結果１１１は、分類結果確認部１１２に表示される。ユーザは、この分類結果が正しいか否かを判断し、もしこの分類結果が間違ったものだと判断した場合には、誤分類の原因となった単語を判定して入力する。すると、分類用辞書学習部１１３が、この誤分類原因単語のデータ（見出し文字列、品詞、分野名）を分類用辞書１０７から削除する。この装置では、このような構成により、学習に用いた分類済み文書群の偏りによるキーワード選択の不備を補正できるようにしている。

【０００８】また、別の従来装置として、特開平７－１１４５７２号公報に示される装置がある。この装置では、文書中の単語の共起関係に基づき単語の特徴ベクトルを生成し、その単語特徴ベクトルから文書の特徴ベクトルを作成し、文書の特徴ベクトル同士の間の類似度を利用して文書を分類する。

【０００９】図３４は、この従来装置の構成図である。

以下、学習時の処理と分類時の処理とに分けて、この従来装置の構成及び機能について説明する。

【００１０】まず学習処理では、まず文書解析部１２２が、文書記憶部１２１内の分類先が既に定まっている学習用文書を取り出し、この文書を解析して出現単語を抽出する。文書解析部１２２で全ての学習用文書から単語の抽出が完了すると、単語ベクトル生成部１２３は、各単語について、その単語と同一文書中に現れる単語（「共起単語」と呼ぶ）を集計し、共起単語を基底としその頻度を成分値とする単語ベクトルを生成する。この単語ベクトルは、単語についての共起単語の傾向を表すベクトルである。得られた単語ベクトルは、単語ベクトル記憶部１２４に記憶される。なお、ここで、文書解析部１２２、単語ベクトル生成部１２３にて処理対象となる単語は、単語ベクトル生成用辞書１２９に登録されている単語に限定される。

【００１１】図３５は、このようにして得られた単語ベクトルの例を示すものである。単語ベクトルの各成分の上に記されている単語が、それら各成分の基底である。図３５では、例えば単語「アメリカ」という単語に対しては、単語「政府」、「先進」、「主要」、「国」、「ココム」が同一文書中に現れた（共起した）ことがあり、共起回数は各１回ずつであったことを示している。

【００１２】対象となる全単語に対する特徴ベクトルの単語ベクトル記憶部１２４への記憶が完了すると、文書ベクトル生成部１２５は、文書解析部１２２で一つの文書から抽出された各単語について、対応する単語ベクトルを当該単語の出現頻度で重み付けして加算することにより、当該文書の特徴を表す文書ベクトルを生成する。例えば、単語「アメリカ」、「兵器」がそれぞれ１回ずつ現れた文書の文書ベクトルは、図３６に示すような形となる。求められた文書ベクトルは、文書ベクトル記憶部１２６に記憶される。

【００１３】次に、文書ベクトル記憶部１２６に記憶された各学習用文書の文書ベクトルに基づき、各分野の特徴を示す代表ベクトルを生成する。ある分野の代表ベクトルは、同一分類に含まれる各学習用文書の文書ベクトルを加算することにより求められる。求められた代表ベクトルは、後述の分類処理において類似度を計算するときにベクトルの大きさが影響しないよう長さ１のベクトルに正規化される。以上で学習処理が終了する。

【００１４】次に分類処理について説明する。まず、文書解析部１２２は、指示された分類対象文書を文書記憶部１２１から取り出し、文書解析を行って出現単語を抽出する。次に、文書ベクトル生成部１２５は、各出現単語に対応する単語ベクトルを出現頻度で重み付けして加算し、当該分類対象文書の文書ベクトルを生成して文書ベクトル記憶部１２６に記憶する。そして、分類部１２７が、この分類対象文書の文書ベクトルに最も類似した代表ベクトルを例えばベクトルの内積演算によって求め、この最類似代表ベクトルに対応する分野に当該分類対象文書を分類する。この分類結果は、分類結果記憶部１２８に記憶される。

【００１５】また、さらに別の従来装置として、『意味属性の学習結果にもとづく文書自動分類方式』（河合、情報処理学会論文誌Ｖｏｌ．３３，Ｎｏ．９，ｐｐ．１１１４―１１２２）に示される装置がある。

【００１６】以下、図３７を参照して、この従来装置について説明する。

【００１７】まず、学習段階では、名詞抽出処理部１３３は、学習用文書１３１を取り出し、単語辞書１４４の情報を用いて、当該学習用文書から名詞を抽出する。次に、意味属性抽出処理部１３６は、学習用文書名詞抽出結果１３４に格納された各名詞の意味属性を、シソーラス１４３から抽出する。シソーラス１４３に格納されている名詞と意味属性の関係の例を図３８に示す。図３８には、例えば「醤油」という名詞が［調味料］という意味属性を有していることが示されている。各名詞の意味属性の抽出結果は、学習用文書意味属性抽出結果１３７に格納される。次に、分類用辞書作成部１３９は、学習用文書名詞抽出結果１３４における各名詞の頻度を分野毎に集計するとともに、学習用文書意味属性抽出結果１３７における各意味属性の頻度を分野毎に集計する。図３９は、各意味属性の各分野毎の頻度の集計結果のテーブルの一例であり、例えば［人工物］という意味属性を持つ単語が「運輸通信」の分野に属する文書に１２回出現したことを示している。単語の集計結果についても同様のテーブルが作成される。分類用辞書作成部１３９は、これら各集計結果のテーブルに対し、統計学におけるカイ２乗検定の考え方を応用した計算式を適用することにより、各名詞と各分野と関連度合いを表したテーブル、及び各意味属性と各分野との関連度合いを表したテーブルを作成する。作成された各テーブルは、分類用辞書１４０に格納される。以上で学習処理が終了する。

【００１８】次に分類時の処理について説明する。まず、名詞抽出処理部１３３は、分類用文書１３２から分類対象に指定された文書を取り出し、単語辞書１４４の情報を用いて、当該文書から名詞を抽出し、分類用文書名詞抽出結果１３５に格納する。次に、意味属性抽出処理部１３６は、抽出された各名詞の意味属性をシソーラス１４３から抽出し、分類用文書意味属性抽出結果１３８に格納する。そして、文書分類部１４１は、まず分類用文書名詞抽出結果１３５と、分類用辞書１４０の単語と分野の関連度合いのテーブルとに基づき、単語の出現頻度からみた当該文書の各分野への関連度合いを計算する。また、文書分類部１４１は、分類用文書意味属性抽出結果１３８と、分類用辞書１４０の意味属性と分野との関連度合いのテーブルとに基づき、意味属性の頻度から見た当該文書の各分野への関連度合いを計算する。そして、文書分類部１４１は、両計算結果を所定の比率で

5

加算することにより、各分野ごとに、当該分類対象文書と当該分類の関連度合いを求める。そして、例えばこの関連度合いの値の最も大きい分野が当該分類対象文書の分類先に選ばれ、文書分類結果１４２に格納される。以上で自動分類処理が終了する。

【００１９】

【発明が解決しようとする課題】以上に説明した従来の各文書自動分類装置は、いずれも、分類済み文書において各分野に特徴的に出現する単語を学習し、この学習結果を分類基準として文書を分類する点では一致する。例えば、特開平７－１１４５７２号公報に示された技術では、ある分野に特徴的に現れる単語は、当該分野の代表ベクトルにおいて大きな値の成分となるので、類似度の値に大きな影響を与え、分類先の決定に大きな影響を与える。また、河合の論文に示された技術でも、ある分野に特徴的に現れる単語は、その分野との関連度合いの値が大きくなるので、分類先を大きく左右する。

【００２０】このような手法は、例えば＜政治＞、＜経済＞などのようにある程度関連が強い単語が共通して出現する分野同士の間では、分類を誤る可能性が高い。例えば、「首相」という単語は、＜政治＞の分野の文書（例えば新聞記事）によく出現する（すなわち特徴的な単語である）が、＜経済＞の分野の文書にもある程度出現する。ここで、＜経済＞分野の文書にたまたま「首相」という単語が多く含まれていると、その文書は「首相」という単語の影響で＜政治＞分野に誤分類されてしまう可能性が高い。

【００２１】このように、従来の文書自動分類技術では、複数の分野に対してそれぞれある程度強い関連を有する単語が分類対象の文書内に数多く現れると、その文書はその単語に対する関連が最も強い分野に分類されやすく、このため誤分類が生じる可能性が高かった。

【００２２】このような傾向は、例えば、＜政治＞、＜科学＞、＜スポーツ＞などのように、出現する単語の傾向の相違が大きい分野同士の間での大まかな分類では致命的な問題にはならないかも知れない。しかしながら、分類を細かくしようとすると、類似する分野が増えてくるので、複数の分野が共通の単語にある程度以上の関連を有するような場合がどうしても増えてくる。このため、従来の文書分類技術は、分類が細かくなると誤分類の可能性が増し、分類精度が劣化するという問題があった。

【００２３】本発明は、このような問題点を解決するためになされたものであり、類似する分野間での分類の精度を向上させることにより、細かい分類に対しても精度のよい分類を行うことができる文書自動分類方法を提供することを特徴とする。

【００２４】

【課題を解決するための手段】この発明は、上記課題を解決するためになされたものであり、分類済みの各文書

6

に出現する各単語の頻度集計結果に基づき各単語と各分野との関連度を登録した関連度テーブルを作成し、この関連度テーブルから、閾値より高い関連度を有する強関連分野が複数存在する複数分野単語を求め、前記関連度テーブルにおける複数分野語についての欄を、当該複数分野語とこれに対応する強関連分野との組合せごとに複数の欄に分割して詳細化関連度テーブルを作成する分類学習ステップと、分類対象の文書に出現する単語の頻度を集計し、この結果得られた頻度情報を前記複数分野語の情報によって詳細化し、この詳細化された頻度情報と前記詳細化関連度テーブルとに基づき当該文書の分類先の分野を決定する分類実行ステップとを含むものである。

【００２５】また、単語の強関連分野の判定基準となる閾値を、関連度テーブルにおける当該単語の各分野に対する関連度の中の最大値に基づき定めるものである。

【００２６】また、分類学習ステップでは、各複数分野語について、当該複数分野語の各強関連分野ごとに、その強関連分野に属する分類済み文書において当該複数分野語と共起した単語の傾向を表す共起ベクトルを生成し、分類実行ステップでは、各複数分野語ごとに、この分類対象文書において当該複数分野語と共起した単語の傾向を示す文書共起ベクトルを生成し、この文書共起ベクトルと前記各共起ベクトルとの類似性に基づき、分類対象文書から得られた前記頻度情報を詳細化するものである。

【００２７】また、各単語の概念的な階層関係を記述したシソーラスを利用して、各共起単語の上位概念の情報を反映した共起ベクトル及び文書共起ベクトルを生成するものである。

【００２８】

【発明の実施の形態】

実施の形態１．以下、この発明の実施の形態を図面を参照して説明する。

【００２９】図１は、この発明に係る文書自動分類方法を実施するためのシステムの構成図である。図１において、分類済み文書記憶部１は、分類済み文書の文書データと、それら各分類済み文書の分類先分野が登録された分類先リストと、を記憶している。これら分類済み文書や分類先リストは、分類基準の学習のために用いられる。また、分類対象文書記憶部２は、自動分類の対象となる分類対象文書の文書データを記憶している。

【００３０】単語分割／頻度抽出部３は、分類済み文書記憶部１又は分類対象文書記憶部２から供給される文書に対し例えば形態素解析などを行うことにより、その文書を単語に分割し、これら単語の頻度統計をとる。そして、単語分割／頻度抽出部３は、その文書内でのそれら各単語の出現位置などの情報を含んだ単語分割情報、及びその文書内での各単語の出現頻度を示す頻度情報を作成する。分類済み文書についての単語分割情報及び頻度

情報（単語分割／頻度情報５１）は、分類済み文書単語分割／頻度情報記憶部４に記憶される。一方、分類対象文書についての単語分割情報及び頻度情報（単語分割／頻度情報５８）は、分類対象文書単語分割／頻度情報記憶部５に記憶される。

【００３１】関連度演算部６は、分類の基準を学習する際において、分類済み文書記憶部１に記憶された各分類済み文書の文書分類先テーブル５２の情報と、分類済み文書単語分割／頻度情報記憶部４に記憶された頻度情報とに基づき、各分野ごとに出現単語の頻度を集計し、この頻度集計の結果に基づき各単語と各分野との関連度を計算する。頻度集計結果及び各単語と各分野との関連度は、それぞれ頻度集計テーブル５３及び関連度テーブル５４として関連度情報記憶部７に記憶される。

【００３２】複数分野語処理部８は、分類学習時には、関連度情報記憶部７の関連度テーブル５４に基づき、関連度の高い分野（「強関連分野」と呼ぶ）が複数存在する単語（「複数分野語」と呼ぶ）を検出し、複数分野語リスト５５を作成する。また、複数分野語処理部８は、この複数分野語リスト５５の情報を用い、分類済み文書単語分割／頻度情報記憶部４の頻度集計テーブル５３を詳細化し、この詳細化された頻度集計テーブルに基づき詳細化関連度テーブル５６を作成する。すなわち、ここでは、各複数分野語を（単語，強関連分野）の組合せごとに別々の単語と捉え直し、単なる単語と分野との関連度だけでなく、複数分野語については（単語，強関連分野）の組合せと各分野との関連度をも含んだ詳細化関連度テーブル５６を作成する。また、複数分野語処理部８は、複数分野語リスト５５と分類済み文書単語分割／頻度情報記憶部４の単語分割情報とに基づき、各複数分野語と例えば同一文書内や同一段落内に現れる単語（共起単語）を求め、これら共起単語の出現傾向を表す共起ベクトル５７を（複数分野語，強関連分野）の各組合せごとについて作成する。

【００３３】このようにして求められた複数分野語リスト５５、詳細化関連度テーブル５６及び各共起ベクトル５７は、分類用情報記憶部９に格納される。この分類用情報記憶部９に格納された情報が、文書の分類先を決定するための基準となる。

【００３４】また、複数分野語処理部８は、文書分類時には、分類対象文書に含まれる各複数分野語について、例えば同一段落における共起単語を検出し、当該分類対象文書における共起単語の傾向を表す文書共起ベクトル５９を作成する。

【００３５】分類先決定部１０は、分類対象文書の出現単語の頻度情報及び文書共起ベクトル５９と分類用情報記憶部９に記憶された各共起ベクトル５７とに基づき、当該文書の分類上の特徴を表す文書ベクトル６０を作成する。そして、分類先決定部１０は、この文書ベクトル６０と分類用情報記憶部９に記憶された詳細化関連度テーブル５６とに基づき、当該文書と各分野との関連度を計算し、この関連度の値に基づき当該文書の分類先の分野を決定する。この結果得られた文書分類結果６１は、文書分類結果記憶部１１に格納される。

【００３６】次に、図２を参照して、この実施の形態のシステムの処理手順の全体的な流れを説明する。図２に示すように、このシステムの処理手順は、分類の基準となる情報を学習する分類学習時の処理と、与えられた分類対象文書を分類する分類実行時の処理とに分かれる。

【００３７】分類学習時においては、まずＳ１にて、単語分割／頻度抽出部３により、分類済み文書記憶部１の各分類済み文書について、出現単語の情報（すなわち単語分割／頻度情報５１）を収集する。Ｓ２では、これらの情報を用いて、関連度演算部６により各単語と各分類との関連度を保持する関連度テーブル５４を作成する。Ｓ３では、複数分野語処理部８が、この関連度テーブル５４から複数分野語を検出し、複数分野語リスト５５を作成する。そして、Ｓ４では、この複数分野語リスト５５に基づき、複数分野語処理部８が、分類の基準となる情報として、詳細化関連度テーブル５６や共起ベクトル５７など、複数分野語を考慮した分類用情報を作成する。

【００３８】そして、個々の未分類の文書の分類を実行する際には、まずＳ５にて、単語分割／頻度抽出処理部３が、分類対象文書の出現単語の情報（すなわち単語分割／頻度情報５８）を収集するとともに、複数分野語処理部８が、単語分割／頻度情報５８に基づき各複数分野語について文書共起ベクトル５９を作成する。次に、Ｓ６では、分類先決定部１０が、単語分割／頻度情報５８、文書共起ベクトル５９、及び分類用情報記憶部９内の各共起ベクトル５７を用いて、当該分類対象文書の分類上の特徴を表す文書ベクトル６０を作成する。そして、Ｓ７では、分類先決定部１０が、文書ベクトル６０と詳細化関連度テーブル５６とに基づき、当該分類対象文書の分類先の分野を決定する。

【００３９】以下、図２の各ステップの処理を更に詳細に説明する。

【００４０】まず図３は、単語分割／頻度抽出部３の処理手順の具体例を示すフローチャートである。図２のＳ１（分類済み文書からの出現単語情報の収集処理）では、単語分割／頻度抽出部３が、この図３の処理手順に従って、各分類済み文書についての単語分割情報及び頻度情報を収集する。以下、図３の処理手順を詳説する。

【００４１】単語分割／頻度抽出部３は、分類済み文書記憶部１から分類済み文書を読み込むと、まずＳ１１において当該文書から１段落を切り出し、この段落の情報を保持する。次に、Ｓ１２にて、切り出した段落から１文を切り出し、この文の情報を保持する。次に、Ｓ１３にて、この文に対して形態素解析を行うことにより、この文から順次単語を切り出す。Ｓ１４では、分割された

9

単語の文書中の位置と品詞の情報を求め、これら情報を分類済み文書単語分割／頻度情報記憶部４内の当該分類済み文書の単語分割情報に登録する。そして、Ｓ１５では、分類済み文書単語分割／頻度情報記憶部４内の当該分類済み文書の頻度情報において、Ｓ１３にて切り出された単語の頻度を１だけ増やし、頻度情報を更新する。なお、Ｓ１４とＳ１５は、いずれを先に行ってもよい。

【００４２】Ｓ１４及びＳ１５の処理は、単語の切り出しが文の末尾に達するまで繰り返される。文の末尾に達すると、Ｓ１２に戻って、段落から次の文を切り出す。また、文の切り出しが段落の末尾に達すると、Ｓ１１に戻って、分類済み文書から次の段落を切り出す。このようにして、分類済み文書の末尾に達するまでＳ１１、Ｓ１２、Ｓ１３、Ｓ１４及びＳ１５の処理が繰り返される。この結果、当該分類済み文書についての単語分割情報及び頻度情報が完成する。この図３の処理は分類済み文書記憶部１に格納された各分類済み文書ごとに繰り返される。

【００４３】この図３の処理を、具体例を用いて説明する。

【００４４】例えば、単語分割／頻度抽出部３に、分類済み文書として図９に示す文書が与えられたとする。この場合、単語分割／頻度抽出部３は、まずＳ１１にて最初の段落７０を切り出し、分類済み文書単語分割／頻度情報記憶部４に作成した当該文書の単語分割情報に、段落の先頭を示す情報を登録する。

【００４５】図１２は、図９の文書から作成された単語分割情報の一例を示す図である。この例では、情報タイプとしてＰ，Ｗ，Ｓの３つのコードが設けられている。コードＰは段落の先頭を示すコードであり、あるＰとその次のＰとに挟まれた部分が、一つの段落についての情報となる。また、コードＳは文の先頭を示すコードであり、あるＳとその次のＳとに挟まれた部分が、一つの文についての情報である。そして、コードＷは単語を示すコードであり、情報タイプがＷの欄には、当該単語の文書冒頭からの位置（バイト単位で示される）と、当該単語を表す文字列と、当該単語の品詞が登録される。図１２は、例えば、図９の文書の最初の段落の最初の文において、文書冒頭から３バイト目の位置に、「内閣」という名詞が出現していることを示している。

【００４６】したがって、Ｓ１１で段落が切り出されると、単語分割情報には情報タイプにコードＰが登録される。

【００４７】次に、単語分割／頻度抽出部３は、Ｓ１２にて段落７０から１文を切り出し、単語分割情報に文の先頭を表すコードＳを情報タイプとして登録する。図９では、段落７０は１文のみしか含まないので、この段落７０（＝１文）について、Ｓ１３以下の処理が行われる。すなわち、Ｓ１３では、従来公知の手法である形態素解析を使ってこの１文を解析し、文中に含まれる単語

10

に分割する。Ｓ１４では、文頭から順番にまず単語「内閣」を取り出す。そして、Ｓ１４では、切り出した単語「内閣」について、文書冒頭からの位置と品詞を求める。この結果、単語分割情報には、図１２に示すように、情報タイプ「Ｗ」、単語位置「３」、単語「内閣」及び品詞「名詞」が登録される。

【００４８】そして、Ｓ１５では、単語分割／頻度抽出部３は、分類済み文書単語分割／頻度情報記憶部４に作成した当該文書の頻度情報に対し、Ｓ１３で分割した単語「内閣」を反映させる。すなわち、単語「内閣」は初出なので、頻度情報に単語「内閣」の欄を作成し、その頻度を１と設定する。なお、初出でない単語の場合は、頻度情報における当該単語の欄の頻度に１を加える。

【００４９】このようにして単語「内閣」についての処理が終わると、段落７０の文から次の単語「支持率」が取り出され、Ｓ１４及びＳ１５でこの単語の情報が単語分割情報及び頻度情報に反映される。

【００５０】このようにして、Ｓ１３で分割した文末までの各単語についてＳ１４及びＳ１５の処理が終了すると、Ｓ１２に戻って段落７０から次の文書を切り出そうとするが、この例では段落７０は１文しか含まないので、Ｓ１１に戻り、分類済み文書から次の段落を切り出し、以上の処理を繰り返す。そして、このような処理を、当該分類済み文書の末尾まで繰り返すことにより、図１１に示す頻度情報と、図１２に示す単語分割情報が得られる。

【００５１】図１１に示す頻度情報では、各単語の頻度についての情報が保持されている。例えば、図１１は、図９の分類済み文書に、「首相」という名詞が３回出現し、「会見」という「サ変名詞」（語尾に「する」を付加することにより動詞として用いることができる名詞）が１回出現したことを示している。なお、以下では、必要に応じ、単語を「単語名：品詞」という形式で表現する。

【００５２】このようにして、分類済み文書記憶部１に格納された全ての分類済み文書に対して図３に示す処理を行うことにより、各分類済み文書の単語分割情報及び頻度情報が得られる。これらの情報は、分類済み文書単語分割／頻度情報記憶部４に記憶される。

【００５３】次に、図２のＳ２（関連度テーブル作成）の処理について、図４を参照して説明する。前述したように、このＳ２の処理は、関連度演算部６によって実行される。図４は、関連度演算部６が実行する処理の手順を示したものである。

【００５４】まず、関連度演算部６は、Ｓ２１にて、分類済み文書記憶部１の文書分類先テーブル５２から分類済み文書を１つ選択し、この文書の分類先を取得すると共に、分類済み文書単語分割／頻度情報記憶部４からこの文書の頻度情報（例えば図１１参照）を取り出す。例えば、図１０は文書分類先テーブル５２のデータ内容の

11

一例を示している。図１０に示すように、文書分類先テーブル５２には、分類済み文書記憶部１に記憶された各分類済み文書ごとに、その文書の文書名（例えばファイル名）と分類先の分野名とが登録されている。一方、分類済み文書単語分割／頻度情報記憶部４には、各分類済み文書の頻度情報及び単語分割情報が、その文書の文書名に対応づけて格納されている。関連度演算部６は、この文書分類先テーブル５２のエントリ（すなわち分類済み文書）を先頭から順次選択し、選択した文書の頻度情報を文書名で検索する。

【００５５】次に、Ｓ２２では、関連度演算部６は、Ｓ２１で取り出した頻度情報に基づき、頻度集計テーブル５３における当該分類済み文書の分類先の各分野について、当該頻度情報に登録された各単語の頻度値をカウントアップする。頻度集計テーブル５３は、例えば、各分野を横軸とし各単語を縦軸とするテーブルである。なお、分類済み文書の分類先が複数分野にわたる場合には、このカウントアップ処理においては、頻度情報における各単語の頻度値をその分類先の分野の数で割った値を頻度集計テーブル５３の該当欄に加える。

【００５６】以上のステップを具体例を用いて説明する。Ｓ２１で例えば図９の分類済み文書が選択され、この文書の文書名が『１１／０４Ｍ／０４－－０９』であったとする。この場合、関連度演算部６は、図１０の分類先リストからこの文書に対応する分野＜政治＞（以下、分野名は＜　＞で括って表現する）を検出すると共に、図１１に示された頻度情報を分類済み文書単語分割／頻度情報記憶部４から取り出す。次に、Ｓ２２では、関連度演算部６は、図１１の頻度情報から、「会見：サ変名詞」の頻度値が１、「首相：名詞」の頻度値が３、などの各単語の頻度値を取得する。この例では、処理対象の文書の分類先は唯一＜政治＞のみであるので、それら各単語の頻度値は、そのまま頻度集計テーブル５３の分野＜政治＞の列における各単語の頻度に加えられる。なお、処理対象が図１０の文書『１１／１２Ｍ／０９－－０８』である場合には、この文書の分類先分野は２つあるので、各単語の頻度を２で除したものが、頻度集計テーブル５３のそれら２分野における各単語の頻度にそれぞれ加えられる。

【００５７】すべての分類済み文書についてＳ２１及びＳ２２の処理が終了すると、頻度集計テーブル５３が完成する。完成した頻度集計テーブル５３のデータ内容の一例を図１３に示す。図１３は、例えば、分類済み文書群の中の＜政治＞分野に分類される文書において「首相：名詞」という単語が５０回出現したことを示している。この頻度集計テーブル５３は、関連度情報記憶部７に格納される。

【００５８】頻度集計テーブル５３が完成すると、次に関連度演算部６は、Ｓ２３にて、このテーブルにおける各単語各分野についての頻度集計結果から、以下に示す

12

式（１）を用いて各単語各分野の理論頻度M_{ij}を算出する。

【００５９】

【数１】

$$M_{ij}=\frac{\sum_{i=1}^{N}F_{ij}}{\sum_{j=1}^{L}(\sum_{i=1}^{N}F_{ij})}\cdot\sum_{j=1}^{L}F_{ij}\qquad(i=1,\ldots,N,\ j=1,\ldots,L)$$

ここで、Ｎは頻度集計テーブル５３における分野の数、Ｌは同テーブルにおける単語の数を示し、ｉは各分野に付した続き番号、ｊは各単語に付した続き番号を示す。また、F_{ij}は分野ｉの文書における単語ｊの出現頻度を示す。

【００６０】式（１）においては、理論頻度M_{ij}は、（単語ｊの総頻度が全単語の総頻度に占める割合）×（分野ｉにおける各出現単語の総数）の形で定義されている。すなわち、理論頻度M_{ij}は、単語ｊが特定の分野に偏らず各分野に平均的に出現すると仮定した場合において、単語ｊが分野ｉに出現する期待される頻度を意味する。この理論頻度については、前述の河合の論文に説明されている。

【００６１】このようにして分野ｉと単語ｊとの組合せごとに理論頻度M_{ij}が求められると、次に関連度演算部６は、Ｓ２４にて、この理論頻度の計算結果と、この元となった頻度集計テーブル５３とに基づき、分野ｉと単語ｊとの関連度Y_{ij}を計算する。ここで、関連度Y_{ij}は、カイ２乗検定を応用した以下の計算式を用いて計算する。

【００６２】

【数２】

$$Y_{ij}=(F_{ij}-M_{ij})\{F_{ij}-M_{ij}\}/M_{ij}\quad(i=1,\ldots,N,j=1,\ldots,L)$$

関連度Y_{ij}の計算結果は、関連度テーブル５４の形で関連度情報記憶部７に格納される。図１４に、図１３の頻度集計テーブルから求められた関連度テーブル５４のデータ内容を例示する。

【００６３】式（１）及び式（２）を用いて求められた分野ｉと単語ｊとの関連度Y_{ij}は、関連が強い（分野，単語）の組合せでは正の大きい値となり、逆に関連が弱い（分野，単語）の組合せについては負の大きな値となる。また、幾ら分野ｉに対して単語ｊの出現頻度が大きかったとしても、その単語ｊが全分野に平均して多く出現する場合は、分野ｉと単語ｊとの関連度Y_{ij}は小さい値（０に近い値）となる。すなわち、ある分野ｉに特異的によく出現する単語ｊがあったとすると、両者の関連度Y_{ij}は正の大きい値となる。また、単語ｊが分野ｉの文書にほとんど出現しなかったような場合には、両者の関連度Y_{ij}は負の大きい値となる。なお、この関連度の考え方も前述の河合の論文に示されている。

【００６４】関連度テーブル５４が完成すると、次に複数分野語処理部８は、Ｓ３にて、このテーブルに基づき複数分野語を検出する。このＳ３の処理の詳細を図５を参照して説明する。

【００６５】まず、複数分野語処理部８は、Ｓ３１で関連度テーブル５４に欄のある全単語について処理が完了したかを検査する。完了していない場合には、Ｓ３２で関連度テーブル５４から単語を１つ選択する。そして、複数分野語処理部８は、Ｓ３３で、関連度テーブル５４を参照し、この単語について関連の強い分野（すなわち強関連分野）を検出し、強関連分野が複数個あった場合にはＳ３４に移る。Ｓ３４では、複数分野語リスト５５にこの単語についての欄を作成し、この単語に対する複数の強関連分野の分野名をこの欄に登録する。なお、Ｓ３３の判定において、強関連分野が１つしかなかった場合には、この単語については何も行わず、Ｓ３１に戻る。そして、関連度テーブル５４に登録された全単語について以上の処理を繰り返すことにより、強関連分野が複数ある単語と、その単語に対応する強関連分野（複数）とが登録された複数分野語リスト５５が完成する。

【００６６】この処理手順において、ある分野が強関連分野であるか否かは、当該分野の関連度を閾値と比較することにより判定する。閾値より大きい関連度を持つ分野は強関連度分野と判定する。この判定のための閾値は、選択した単語についての各分野の関連度のうちの最大値に所定の係数（１以下の正数）を乗じた値を用いる（したがって、どの単語も最低一つは強関連分野を有する）。

【００６７】すなわち、複数分野語処理部８は、選択した単語ｊについて、関連度テーブル５４におけるｊ番目の行の各分野の関連度を取り出し、これら関連度に次式（３）を適用することにより、当該単語ｊの強関連度分類の数T_jを算出する。

【００６８】

【数３】

$$T_j = \#\{Y_{ij} \mid Y_{ij} > \max_{i=1}^{N}(Y_{ij}) \cdot Y_{border}\} \quad (j = 1, \ldots, L)$$

ここで、演算子『＃』は、後続の集合『｛｝』に含まれる要素（Y_{ij}）の数を返す。また、Y_{border}は、強関連分類の判定のための閾値を求める際の所定の係数（固定値）である。

【００６９】この式（３）によれば、単語ｊの各分野に対する関連度の最大値に対して所定の割合となる値を閾値とし、関連度Y_{ij}がその閾値より大きい分野（すなわち強関連分野）の数がT_jとして求められる。このT_jの値が２以上の場合、単語ｊは複数分野語と判定され、複数分野語リスト５５にその単語名と各強関連分野の分野名が登録される。

【００７０】具体例を用いて説明する。例えば、図１４の関連度テーブルが与えられ、係数Y_{border}の値を０．３にした場合、単語「首相」については、最大の関連度を持つ分野＜政治＞の関連度（６６．７）にこの係数０．３を乗じたものが、この単語「首相」についての強関連分野判定の閾値となる。そして、この単語「首相」に対する分野＜経済＞の関連度（２０．８）はその閾値より大きいので、単語「首相」は複数分野語と判定される。したがって、複数分野語リスト５５には、図１５に示すように、単語「首相：名詞」と共に、これに対応する強関連分野として＜政治＞及び＜経済＞が登録される。このような処理を関連度テーブルにおける全単語について行うことにより、図１５に示すような複数分野語リストが完成する。図１５のリストは、例えば単語「政治：名詞」は分野＜政治＞及び＜経済＞に関連が強く、単語「市場：名詞」に分野＜経済＞及び＜国際＞に関連が強いことを示している。

【００７１】このようにして作成された複数分野語リスト５５は、分類用情報記憶部９に記憶され、未分類文書の分類のために利用される。

【００７２】なお、この実施の形態では、前述のように強関連分類の判定のための閾値を固定値とはせず、各単語の関連度の最大値に合わせて求めることとしたので、各単語の関連度の分布に合わせ、相対的にみて関連度が高い分野を強関連分野として抽出することができる。すなわち、例えば学習に用いた分類済み文書群においてたまたま出現回数が少なかったような単語についても、その単語の強関連分野を適切に求めることができる。

【００７３】このようにして複数分野語リスト５５が完成すると、次に複数分野語処理部８は、Ｓ４にて、それら複数分野語を考慮して、未分類の文書を分類する際の分類基準となる分類用情報を作成する。ここでは、分類用情報として、前述の関連度テーブル５４を複数分野語を考慮して詳細化した詳細化関連度テーブル５６と、各複数分野語についての共起単語の傾向を示す共起ベクトル５７を作成する。以下、詳細化関連度テーブル５６の作成手順及び共起ベクトル５７の作成手順を順に説明する。なお、詳細化関連度テーブル５６の作成と共起ベクトル５７の作成はいずれを先に行ってもよい。

【００７４】まず、図６を用いて詳細化関連度テーブル５６の作成手順を説明する。詳細化関連度テーブル５６は、例えば、図１９に示すように、複数分野語を（単語，強関連分野）の組合せごとに別々の単語と捉えて分割（例えば「首相」を「首相（政治）」と「首相（経済）」に分割）し、関連度を求め直した結果である。このテーブルの作成手順は、大まかに分けて、複数分野語を（単語，強関連分野）の組合せごとに別々の単語とみなして頻度集計テーブル５３（例えば図１３参照）を詳細化するプロセスと、この結果得られた詳細化頻度集計テーブルから詳細化関連テーブルを作成するプロセスと、を含んでいる。図６では、Ｓ４０１～Ｓ４０５が前者のプロセスに対応し、Ｓ４０６～Ｓ４０７が後者のプ

15

ロセスに対応する。

【００７５】すなわち、まず複数分野語処理部８は、Ｓ４０１にて、頻度集計テーブル５３に登録された全単語について処理が完了したか検査する。完了していない場合には、Ｓ４０２にて頻度集計テーブル５３から未処理の単語を1 つ取り出す。そして、Ｓ４０３にて、選択した単語が複数分野語リスト５５に登録された単語であるか否か（すなわち複数分野語か否か）チェックする。その単語が、複数分野語リスト５５に登録されていた場合は、複数分野語処理部８は、Ｓ４０４にて、当該単語（すなわち複数分野語）と、これに対応する複数の強関連分類との各組合せごとに、詳細化頻度集計テーブルに欄を作成し、頻度集計テーブル５３における当該単語の各分野での頻度値を、それら各組合せの欄に配分する。

【００７６】すなわち、図１８に示すように、複数分野語（例えば「首相」）を各強関連分野（「首相」については＜政治＞、＜経済＞）ごとに別々の単語（単語「首相（政治）」及び単語「首相（経済）」）と捉え、これら新たな単語について詳細化頻度集計テーブルに欄を作成する。そして、このようにして作成した欄「単語（分野）」における各分野に対し、頻度集計テーブルの頻度値を配分していく。配分する頻度値は次のように決定する。まず、当該「単語（分野）」の（）内の「分野」に一致する分野については、頻度集計テーブル５３におけるその分野の頻度値をそのまま設定する。また、当該「単語（分野）」の（）内の「分野」には一致しないが、当該「単語」の強関連分野の一つである分野については、その頻度を０とする。そして、当該「単語」の強関連分野以外である分野については、頻度集計テーブル５３におけるその分野についての頻度を当該単語の強関連分野の数で割り、その結果を設定する。この配分の仕方を具体的に示すと次のようになる。

【００７７】図１３の頻度集計テーブルから図１８の詳細化頻度集計テーブルを作成する場合を例にとって説明する。複数分野語処理部８は、図１３のテーブルから「首相」の各分野についての頻度値を取り出し、これらを順番に図１８のテーブルの「首相（政治）」及び「首相（経済）」に配分していく。まず、図１３における「首相」の＜政治＞に対する頻度５０は、＜政治＞は「首相」の強関連分野なので、（）内の分野の一致する「首相（政治）」にそのすべてが配分され、（）内の分野の異なる「首相（経済）」には全く配分されない。したがって、図１８の詳細化頻度集計テーブルの分野＜政治＞においては、「首相（政治）」の頻度は５０、「首相（経済）」の頻度は０と設定される。逆に、同じく「首相」の強関連分野である＜経済＞については、図１３における頻度３０がすべて「首相（経済）」に対して配分され、「首相（政治）」の頻度は０に設定される。また、図１３における「首相」の＜労働＞に対する頻度３は、＜労働＞は「首相」の強関連分野ではないので、

16

これを「首相」の強関連分野の数２で割った結果の１．５が「首相（政治）」と「首相（経済）」とに等しく配分される。したがって、図１８のテーブルの分野＜労働＞においては、「首相（政治）」、「首相（経済）」が共に頻度１．５となる。

【００７８】一方、Ｓ４０３で、選択した単語が複数分野語リスト５５にない単語と判定された場合は、複数分野語処理部８は、Ｓ４０５にて、頻度集計テーブル５３における当該単語の各分野についての頻度を、そのまま詳細化頻度集計テーブルに設定する。

【００７９】例えば、図１３の頻度集計テーブルにおける単語「薄商い」は、複数分野語ではないので、頻度集計テーブルにおける「薄商い」の各分野についての頻度が、そのまま詳細化頻度集計テーブルにおける「薄商い」の各分野についての頻度として設定される。

【００８０】このような処理を頻度集計テーブル５３の全登録単語に対して行うことにより、図１８に示すように、複数分野語を強関連分類語ごとに分割することにより詳細化された詳細化頻度集計テーブルが完成する。

【００８１】詳細化頻度集計テーブルが完成すると、次に複数分野語処理部８は、このテーブルに対して前述の式（１）を適用して理論頻度を算出する（Ｓ４０６）。そして、Ｓ２の場合と同様に、この理論頻度の算出結果を用いてカイ２乗検定の応用である前述の式（２）の演算を行うことにより、詳細化頻度集計テーブルにおける単語、分野の各組合せごとに、それら両者の関連度を算出する。この結果、複数分野語を強関連分類語ごとに別々の単語に分割した場合における、各単語と各分野との関連度を示した詳細化関連度テーブルが完成する。

【００８２】図１９は、このようにして作成された詳細化関連度テーブルの一例である。図１９のテーブルは、図１８の詳細化頻度集計テーブルから作成されたものである。図１９の詳細化関連度テーブルを図１４の関連度テーブルと比較すると、例えば図１４における「首相」の＜政治＞に対する関連度と比較した場合、図１９の「首相（政治）」の＜政治＞に対する関連度はそれより大きくなっており、「首相（経済）」の＜政治＞に対する関連度はそれより小さくなっているのが分かる。逆に、図１４における「首相」の＜経済＞に対する関連度と比較した場合、図１９の「首相（政治）」の＜経済＞に対する関連度はそれより小さくなり、「首相（経済）」の＜経済＞に対する関連度はそれより大きくなっている。

【００８３】このようにして作成された詳細化関連度テーブル５６は、分類用情報記憶部９に格納される。

【００８４】次に共起ベクトル５７の作成手順について説明する。この実施の形態では、すでに説明したように、複数分野語についてのみ共起ベクトルを作成する。しかも、この共起ベクトルは、当該複数分野語の強関連分野を考慮して、複数分野語と強関連分野との組合せご

とに作成する。以下、図７を参照して、共起ベクトルの作成手順を説明する。

【００８５】まず、複数分野語処理部８は、Ｓ４１１にて、全分類済み文書に対して処理が完了したかを検査する。完了していない場合には、Ｓ４１２にて、分類済み文書を一つ選択し、その文書の単語分割情報と頻度情報を分類済み文書単語分割／頻度情報記憶部４から取り出す。次に、複数分野語処理部８は、取り出した頻度情報にあるすべての単語について処理が終わったかをＳ４１３で検査する。終わっていない場合は、Ｓ４１４にて、頻度情報から単語を一つ選択する。そして、Ｓ４１５では、複数分野語処理部８は、選択した単語が複数分野語かどうかを複数分野語リスト５５を参照して判定する。ここで、複数分野語でない場合は、その単語については何もせず、Ｓ４１３に戻る。

【００８６】Ｓ４１５において当該単語が複数分野語であった場合は、複数分野語処理部８は、Ｓ4,１６において、現在処理中の文書の分類先の分野を文書分類先テーブル５２にて調べ、その分野が複数分野語リスト５５における当該単語の強関連分野に含まれているかどうかを判定する。

【００８７】この判定の結果、現在処理中の文書の分類先が、当該単語の強関連分野に含まれていなければ、当該単語については何も処理を行わずにＳ４１３に戻る。一方、Ｓ４１６において、現在処理中の文書の分類先が、当該単語の強関連分野に含まれている場合には、複数分野語処理部８は、Ｓ４１７にて、当該文書の単語分割情報を参照して、例えば同一段落などの所定の範囲内において当該単語と共起する単語を当該文書全体にわたって調べる。この時、これら各共起単語の頻度も同時にカウントする。そして、複数分野語処理部８は、検出した共起単語とその頻度とを、Ｓ４１８にて共起頻度情報に反映させる。このようにして１つの単語についての処理が終わるとＳ４１３に戻って次の単語の処理に移る。

【００８８】このような処理よれば、各複数分野語ごとに、当該複数分野語の強関連分野に属する文書において当該複数分野語と共起した単語及びその頻度の傾向が、共起頻度情報として求められる。

【００８９】図１６は、図１２の単語分割情報及び図１５の複数分野語リストに従って作成された共起頻度情報の内容を模式的に示した図である。図１６では、「首相」、「市場」などの各複数分野語に対して、それぞれ対応する強関連分野が関連付けられており、さらにその複数分野語と強関連分野との組合せに対して、それぞれ共起単語及びその頻度を含んだデータのリストが関連付けられている。図１６は、例えば、「総裁」という単語が、＜政治＞分野の文書では複数分類語「首相」に対して５回共起し、＜経済＞分野の文書では複数分類語「首相」に対して３回共起したことを示している。

【００９０】ここまでの処理の具体例を以下に示す。まず、図９の文書がＳ４１２で選択されたとする。この場合、複数分野語処理部８は、この文書に対応する頻度情報及び単語分割情報として、図１１及び図１２に示した情報を分類済み文書単語分割／頻度情報記憶部４から取り出す。そして、頻度情報の全ての単語について処理が終わるまでＳ４１３以下の処理を繰り返す。例えば、Ｓ４１４で図１１から「会見：サ変名詞」が選択されたときには、この単語は図１５の複数分野語リストに存在しないため、Ｓ４１５の判定により何も行わずにＳ４１３に戻る。順次処理が進んで、Ｓ４１３で「首相：名詞」が選択されたときには、Ｓ４１５でこの単語は図１５の複数分野語リストに存在することを検知し、Ｓ４１６に進む。Ｓ４１６では、図１０の文書分類先テーブルから、図９の文書の分類先が＜政治＞分野であることを検知するとともに、複数分野語リストから当該複数分野語「首相：名詞」の強関連分野が＜政治＞及び＜経済＞であることを検知し、これらのことから、当該文書の分類先の分野が当該複数分野語の強関連分野の一つであることを検知する。したがって、Ｓ４１６の判定結果はＹＥＳとなり、以下、当該文書の分類先分野と当該複数分野語の強関連分野との一致点である＜政治＞分野について、Ｓ４１７及びＳ４１８の処理が行われる。この具体例では、Ｓ４１７にて、現在処理中の複数分野語と同じ段落に出現したものを共起単語として抽出する。したがって、図９の段落７０からは、複数分野語「首相：名詞」の共起単語として、「内閣：名詞」、「支持率：名詞」、「理由：名詞」、「海部：名詞」、「続投論：名詞」、「三塚：名詞」、「反発：サ変名詞」が抽出される。また、これら各共起単語の同段落７０における頻度（いずれも１である）も検出される。そして、Ｓ４１８では、図１６に示す共起頻度情報において、「首相：名詞」の＜政治＞に関連付けられた各共起単語の頻度にそれぞれＳ４１７で検出された頻度を加える。このような処理をすべての単語について繰り返すことにより、図１６に示すような共起頻度情報が得られる。

【００９１】なお、この例では、同一段落に出現したものを共起単語としたが、これに限らず、広く同一文書に出現したものすべてを共起単語としてもよいし、逆に範囲を狭め、同一の文に出現したもののみを共起単語としてもよい。この実施の形態では、単語分割情報（たとえば図１２）を作成しているので、このようないずれの場合にも対応することができる。

【００９２】このようにして共起頻度情報が完成すると、複数分野語処理部８は、この共起頻度情報から、各複数分野語・強関連分野の組合せごとについて、共起ベクトルを作成する。共起ベクトルは、共起頻度情報（例えば図１６）における共起単語の頻度値をその共起単語に対応する各基底の成分値とするベクトルである。ただし、この実施の形態では、学習に用いた分類済み文書群に現れたすべての単語を共起ベクトルの基底とし、共起

しなかった単語に対する成分値は０とすることで共起ベクトルの基底を統一している。例えば、図１６の共起頻度情報から求めた「首相（政治）」、「首相（経済）」の共起ベクトルは、それぞれ図１７の（ａ）に示すようなものとなる。

【００９３】そして、複数分野語処理部８は、Ｓ４１９で、各共起ベクトルを長さ１に正規化することにより、共起ベクトルの長さの差を吸収する。すなわち、以下では、共起ベクトルの「方向」のみについて注目する。この方向が、複数分野語と強関連分野との組合せである「単語（分野）」に対する共起単語の出現傾向を表す。例えば、図１７の（ａ）に示した各共起ベクトルは、同図の（ｂ）のように正規化される。

【００９４】このようにして作成された各複数分野語と強関連分野との各組合せについての共起ベクトルは、分類用情報記憶部９に格納される。

【００９５】これで、分類用情報記憶部９には、分類対象文書の分類処理の際の基準となる複数分野語リスト５５、詳細化関連度テーブル５６及び各共起ベクトル５７がすべて用意された。以下、これらの情報を用いた文書の分類処理の各ステップ（図２のＳ５～Ｓ７）の処理手順を更に詳細に説明する。

【００９６】まず、分類対象文書が与えられた場合、まずＳ５にて、単語分割／頻度抽出処理部３が、図３に示した処理手順により当該分類対象文書の出現単語を解析し、当該分類対象文書の単語分割情報及び頻度情報を作成する。この手順は、分類済み文書の場合と同様なので説明は省略する。

【００９７】例えば、分類対象文書として図２０に示す文書が与えられたとする。この文書は、本来＜経済＞に分類されるべき文書である。単語分割／頻度抽出処理部３は、この文書を解析し、図２１に示す頻度情報及び図２２に示す単語分割情報を作成する。例えば、図２１は、図２０の文書に、「東京証券取引所：名詞」が１回、「首相：名詞」が３回出現していることを示している。また、図２２は、例えば、「首相：名詞」が文書冒頭から２２３バイト目に出現したことを示している。

【００９８】このようにして得られた分類対象文書の分類対象文書の単語分割情報及び頻度情報（単語分割／頻度情報５８）は、分類対象文書単語分割／頻度情報記憶部５に格納される。

【００９９】分類対象文書の単語分割情報及び頻度情報が得られると、次に複数分野語処理部８及び分類先決定部１０により、当該分類対象文書の特徴を表す文書ベクトル６０を作成し（Ｓ６）、この文書ベクトル６０に基づき分類先を決定する（Ｓ７）。この一連の処理の詳細な手順を、図８を参照して説明する。なお、図８における各ステップは、Ｓ６１～Ｓ６６が図２のＳ６に対応し、Ｓ７０が図２のＳ７に対応する。

【０１００】分類先決定部１０は、分類対象文書の頻度

情報（例えば図２１）を取り出し、この頻度情報にある単語を先頭から順に一つずつ取り出して処理していく。このため、まずＳ６１にて、その頻度情報にあるすべての単語に対して処理が終わったかを検査する。終わっていない場合は、Ｓ６２で、頻度情報における次の未処理単語を選択する。そして、Ｓ６３では、分類用情報記憶部９に格納された共起ベクトルの情報から、選択した単語が共起ベクトルを持つ単語かどうかを判定する。なお、この判定は、当該単語が複数分野語リスト５５に含まれるかどうかに基づき行ってもよい。

【０１０１】Ｓ６３の判定の結果、当該単語が共起ベクトルを持たない単語であれば、分類先決定部１０は、頻度情報における当該単語の頻度を、そのまま文書ベクトルにおける当該単語に対応する成分に設定する。

【０１０２】ここで、文書ベクトルは、学習に用いた分類済み文書群に現れる全単語を基底とするベクトルであり、基本的には、分類対象文書に出現した単語の頻度をその単語に対応する基底の成分値とするベクトルである（したがって、出現しなかった単語についての成分は０となる）。このため、Ｓ６６では、分類対象文書における単語の頻度値を文書ベクトルに設定する。例えば、図２１の頻度情報における「東京証券取引所：名詞」という単語は、共起ベクトルを持たないので、その単語の頻度１が、そのまま文書ベクトルにおける「東京証券取引所：名詞」の成分に設定される。

【０１０３】ただし、文書ベクトルでは、複数分野語については各強関連分野との組合せごとに１単語（例えば、分野付きの単語「首相（政治）」など）とみなし、各組合せをそれぞれ基底としている。すなわち、この文書ベクトルは、基本的には分類対象文書における単語の出現傾向を示すベクトルであるが、更に複数分野語がどの分野（強関連分野）の単語として出現したかを示す情報を含んだものとなっている。

【０１０４】このような文書ベクトルを作成のため、分類対象文書内に出現した複数分野語については、その頻度を各強関連分野との組合せごとに分配する必要がある。このための処理がＳ６４及びＳ６５の各ステップである。

【０１０５】すなわち、Ｓ６３の判定で共起ベクトルを持つ単語（すなわち複数分野語）と判定された場合、Ｓ６４では、まず複数分野語処理部８にて、当該単語の文書共起ベクトルを作成する。ここで、文書共起ベクトルは、前述の共起ベクトルと同様、文書の所定範囲内（例えば同一段落内）で当該単語と共起した単語の頻度を各成分値としたベクトルである。共起ベクトルと文書共起ベクトルとの相違は、前者は学習に用いた複数の分類済み文書から作成されたものであるのに対し、後者は１つの分類対象文書のみから作成されたものである点である。すなわち、文書共起ベクトルは、複数分野語が、分類対象文書においてどのような単語と共起したかという

共起単語の傾向を表すベクトルである。

【０１０６】Ｓ６３で共起ベクトルを持つと判定された単語についての文書共起ベクトルの作成は、共起ベクトルの作成手順（図７参照）におけるＳ４１７及びＳ４１８とほぼ同様の処理にて行うことができる。すなわち、単語分割情報（例えば図２２）を参照して当該単語の共起単語を検出し、それら共起単語の頻度をそれぞれ対応する成分の値として設定すればよい。

【０１０７】例えば、図２０の分類対象文書に出現する単語「首相：名詞」は、図１７などに示すように共起ベクトルを持つ複数分野語である。この図２０の文書から作成した「首相：名詞」の文書共起ベクトルを図２３に示す。図２３の文書共起ベクトルは、共起の範囲を同一段落内に限った場合の例である。図２３と図１７を比較すれば分かるように、文書共起ベクトルは、共起ベクトルと同じ基底のベクトルである。なお、この文書共起ベクトルの長さを正規化する必要はない。

【０１０８】選択した単語についての文書共起ベクトルが求められると、Ｓ６４では、更に分類先決定部１０が、分類用情報記憶部９から当該単語についての各共起ベクトルを取り出し、文書共起ベクトルとこれら各共起ベクトルとの内積をそれぞれ計算する。得られた各内積値は、当該単語（この単語は複数の分野に強い関連を持つ複数分野語である）が、当該分類対象文書においてはどの分野の単語として出現している可能性が高いかを示している。

【０１０９】すなわち、各共起ベクトルは、複数分野語と強関連分野との組合せについて求められており、例えば「首相（政治）」についての共起ベクトルは、「首相」という複数分野語が＜政治＞分野の文書に出現する場合にはどのような単語と共起しているかという傾向を表している。したがって、分類対象文書における当該単語の共起単語の傾向を表す文書共起ベクトルと、各分野の文書における当該単語の共起単語の傾向を表す各共起ベクトルとを比較すれば、それらの類似度合いから分類対象文書において当該単語はどの分野の単語として現れているかを知ることができる。

【０１１０】共起ベクトルはすべて長さ１に正規化されているので、文書共起ベクトルと共起ベクトルの内積値の大小は、純粋に両ベクトルの方向の類似性のみを示し、内積値が大きいほど両ベクトルの方向は類似しているといえる。文書共起ベクトルや共起ベクトルの方向は、共起単語の出現の傾向を表しているので、これら両ベクトルの内積値を、共起単語の出現傾向の類似度合いを示していると捉えることができる。したがって、この実施の形態では、文書共起ベクトルと各共起ベクトルの内積値を、当該単語がどの分野の単語であるかを示す指標値として用いる。例えば、図２３の「首相」の文書共起ベクトルと、図１７（ｂ）の「首相（政治）」の共起ベクトルとの内積値は、分類対象文書に出現した単語

「首相」が分野＜政治＞の単語である可能性の大きさを示している。同様に、図１７（ｂ）の「首相（経済）」の共起ベクトルとの内積値は、「首相」が＜経済＞の単語である可能性の大きさを示している。

【０１１１】そこで、分類先決定部１０は、Ｓ６５にて、求められた各内積値の大きさの比率に従って、頻度情報（図２１参照）における当該単語の頻度値を、当該単語と強関連分野との各組合せに対して比例配分し、その配分結果を文書ベクトルに設定する。

【０１１２】例えば、前述の単語「首相」についての例を用いて説明すると、「首相」の文書共起ベクトル（図２３）と、「首相（政治）」「首相（経済）」についての各共起ベクトル（図１７）との内積値は、それぞれ０．８２及び４．０８となるので、分類対象文書における「首相」の頻度３を、それら内積値の比率に応じて比例配分すると、「首相（政治）」に配分される頻度値は、

【数４】

$$\frac{0.82}{(0.82+4.08)} \times 3 = 0.5$$

となり、「首相（経済）」に配分される頻度値は、

【数５】

$$\frac{4.08}{(0.82+4.08)} \times 3 = 2.5$$

となる。したがって、Ｓ６５では、文書ベクトルの「首相（政治）」及び「首相（経済）」の成分値に、０．５及び２．５をそれぞれ設定する。

【０１１３】以上説明したＳ６１～Ｓ６６の処理を、分類対象文書の頻度情報のすべての単語について繰り返すことにより、当該分類対象文書の文書ベクトルが求められる。例えば、図２４は、分類対象文書の頻度情報（図２１）の各単語の頻度を、各単語の共起ベクトル及び文書共起ベクトルを使って分割した結果を示す例である。この表における各単語の頻度を、所定の順番（図１９等に示す詳細化関連度テーブルでの単語の登録順序）に合せて並べたものが、文書ベクトルとなる。このようにして作成された文書ベクトルは、分類対象文書の出現単語の傾向を表している。

【０１１４】分類対象文書の文書ベクトルが完成すると、分類先決定部１０は、Ｓ７０にて、この文書ベクトルと分類用情報記憶部９の詳細化関連度テーブルとを用いて、当該分類対象文書の分類先を決定する。このため、分類先決定部１０は、次の式を用いて当該分類対象文書の各分野ｉへの関連度S_iを計算する。

【０１１５】

【数６】

$$S_i = \sum_{j=1}^{L} (Y_{ij} \cdot D_j) \quad (i = 1, \ldots, N)$$

この式で、D_jは文書ベクトルにおける単語ｊに対応す

る成分値を示し、Ｌ´は複数分野語を各強関連分野ごとに別々の単語（例えば「首相（政治）」、「首相（経済）」）とみなしたとき単語の総数を示す。この演算は、詳細化関連度テーブル（例えば図１９）を各分野ごとに分割し、各分野の列をそれぞれ当該分野の単語の出現傾向を示すベクトルとみなし、このベクトルと文書ベクトルとの内積を求める演算と捉えることができる。

【０１１６】例えば、図２０の文書の各分野に対する関連度は、図１９の詳細化関連度テーブルと図２４の表に対応する文書ベクトルとに基づき、図２５に示すような内積演算で求められる。図２５に示すように、図２０の文書は、分野＜経済＞に対する関連度が最も大きくなっている。

【０１１７】分類先決定部１０は、例えば関連度S_iが最大となる分野を、分類対象文書の分類先に決定する。なお、分類先を一つに限らずに、例えば関連度S_iが所定順位内の分野を分類先としてもよい。また、関連度S_iが最大となる分野を分類先として抽出するのに加え、その関連度の最大値に対して所定の割合以上の関連度を持つ分野を副分類先として抽出するような応用も考えられる。得られた文書分類結果６１は、文書分類結果記憶部１１に格納される。

【０１１８】図２５の例では、図２０の文書が分野＜経済＞に正しく分類されていることが分かる。

【０１１９】これに対し、図２６は、図２０の文書の分類先を、複数分野語を分割していない図１４の関連度テーブル及び図２１の頻度情報から求めた場合の演算例を示している。図２６の例では、分野＜政治＞に対する関連度が最大値となっており、これでは図２０の文書が、本来分類されるべき＜経済＞ではなく、＜政治＞に誤って分類されてしまう。このように、複数分野語に注目しない従来の手法では、分類先を誤る可能性が高いことが分かる。これに対し、この実施の形態の手法は、複数分野語に注目して、複数分野語を各強関連分野ごとに別々の単語に分割して取り扱うことにより、文書における単語の出現傾向をより詳細に分析することができるので、より好ましい分類結果を得ることができる。

【０１２０】以上説明したように、この実施の形態１によれば、分類済み文書を解析することにより複数分野語を抽出し、この複数分野語に注目して、分類の際の基準となる関連度テーブルや、分類対象文書の出現単語の頻度情報を詳細化することにより、分類対象文書の各分野への関連度をより詳細に分析することができるので、類似する分野（共通の単語がある程度以上の頻度で出現する複数の分野）間での分類の精度を向上させることができる。したがって、この実施の形態１は、分類を細かくして類似する分野が多くなるような場合においても、分類精度の劣化を抑えることができる。

【０１２１】また、この実施の形態１は、これら関連度テーブルなどの詳細化を、複数分野語についての共起単

語の傾向を示すベクトル（共起ベクトル、文書共起ベクトル）に基づき行うことにより、複数分野語が現れた周囲の状況に基づき当該複数分野語がどの分野の単語として現れたかを評価することができる。これにより、従来技術において誤分類の原因となっていた複数分野語を、各強関連分野ごとに別々の単語として切り分けて取り扱うことが可能となるとともに、複数分野語の出現状況を踏まえて適切にその切り分けを行うことができる。

【０１２２】実施の形態２．次に、この発明の実施の形態２を説明する。この実施の形態２は、各単語の上位概念語が登録されたシソーラスを利用して、共起ベクトルや文書共起ベクトルを上位概念語を含めた形式に拡張することにより、複数分野語の頻度の分割をより適切に行おうとするものである。

【０１２３】この実施の形態２のシステムの構成を図２７に示す。図２７において、図１と同様の構成要素には同一の符号を付してその説明を省略する。図２７から分かるように、この実施の形態２の各構成要素は、シソーラス利用型複数分野語処理部２１とシソーラス２２以外は、図１の構成要素と同一である。シソーラス２２には、各単語の意味概念の階層関係の情報を格納している。例えばシソーラス２２は、単語「テニス」の上位概念として単語「球技」があり、単語「球技」の上位概念として単語「スポーツ」があるなど、各単語の階層的な上下関係の情報を記憶している。シソーラス利用型複数分野語処理部２１は、図１（実施の形態１）における複数分野語処理部８に対応するものであり、基本的な処理動作は共通している。ただし、シソーラス利用型複数分野語処理部２１は、共起ベクトルや文書共起ベクトルを作成する際に、このシソーラス２２を利用して各共起単語の上位概念語を抽出し、この上位概念語をそれらベクトルに反映させる点が、図１の複数分野語処理部８と異なる。

【０１２４】この実施の形態２のシステムは、シソーラス利用型複数分野語処理部２１における共起ベクトル及び文書共起ベクトルの作成手順のみが実施の形態１のシステムと異なるだけで、その他の処理手順は、実施の形態１の場合と全く同じでよい。したがって、以下では、実施の形態１と異なる部分である、シソーラス利用型複数分野語処理部２１における共起ベクトルの作成手順のみを詳細に説明する。

【０１２５】図２８は、このシソーラスを利用した分類学習時の共起ベクトルの作成手順を示すフローチャートである。図８（実施の形態１での共起ベクトル作成処理）におけるステップと同じ処理を示すステップについては、図２５においても図８と同じ符号を付してその説明を省略する。

【０１２６】図２８において、分類済み文書の頻度情報から複数分野語を取り出し、この複数分野語の共起単語を単語分割情報から抽出するまで（Ｓ４１７まで）の処

理は、実施の形態１と同様の処理である。このようにして、複数分野語の共起単語が取り出されると、シソーラス利用型複数分野語処理部２１は、各共起単語の上位概念語をシソーラス２２から取り出す。取り出した上位概念語には、共起単語の頻度に対し階層差に応じた重みを乗じた値を頻度値として割り当てる。そして、Ｓ５０２では、共起頻度情報（例えば図１６参照）に対し、各共起単語の頻度値を反映させる。このとき、共起単語だけでなく、それらの上位概念語も共起頻度情報のリストに加え、その頻度値を反映させる。

【０１２７】例えば、Ｓ４１７にて、共起単語として「テニス」が抽出され、シソーラス２２には図２９に示すような単語の階層関係が記憶されていたとする。この場合、Ｓ５０１では、「テニス」の上位概念語として「球技」及び「スポーツ」が抽出される。ここでは、上位概念語には、例えば、共起単語から階層が１つ上に上がるごとに所定の重みを乗じた頻度値を割り当てる。例えば、共起単語「テニス」の頻度が１であり、前記所定の重みを０．５とすると、Ｓ５０１では、「テニス」「球技」「スポーツ」の各単語に対し、図２９に示すようにそれぞれ１，０．５，０．２５の頻度値が割り当てられる。そして、これら頻度値が、Ｓ５０２で共起頻度情報における対応単語の頻度値に加算される。そして、この共起頻度情報から共起ベクトルが作成され、長さ１に正規化される。

【０１２８】また、この実施の形態２では、文書共起ベクトルにも、上記共起ベクトルの場合と同様に、共起単語の上位概念語を反映させる。

【０１２９】すなわち、前記実施の形態１では、図３０の（ａ）に示すように共起ベクトルには共起単語「テニス」そのものの頻度しか反映させていなかったが、この実施の形態２では、共起ベクトルに対し共起単語の上位概念語「球技」及び「スポーツ」の頻度も反映させる（なお、図３０では、共起ベクトルの正規化は行っていない）。

【０１３０】分類済み文書の数は有限なので、分類学習時において、ある複数分野語の共起単語として現れる単語もある程度限られてくる。したがって、これから分類しようとする分類対象文書に、そのような限られた単語そのものが当該複数分野語の共起単語として現れるとは限らない。例えば、分類対象文書に、当該複数分野語の共起単語として、分類学習時の共起単語そのものは現れなかったが、それに類似した単語は現れたというような場合も考えられる（実際に、類似した単語が現れる確率は高いと考えられる）。このような場合、分類学習時に実際に共起した単語のみしか共起ベクトルに考慮しないとすれば、分類対象文書における複数分野語の頻度をある強関連分野に配分する際に、たまたまその分類学習時における当該強関連分野の共起単語そのものが現れなかっただけのために、その強関連分野に対する配分比率が

下がってしまうというようなこともあり得る（類似の単語は出現しているのだから、実際にはその複数分野語はその強関連分野に対してもっと関連が深いはずである）。このような誤りが生じるのは、分類対象文書にせっかく現れた類似単語の情報が捨てられてしまうためといえる。

【０１３１】これに対し、この実施の形態２では、共起単語そのものだけでなく、その類似単語の一種である上位概念語をも共起ベクトル、文書共起ベクトルに反映させるので、分類対象文書内にある複数分野語に対する分類学習時の共起単語そのものが現れない場合でも、その上位概念語が共起単語として現れていれば、当該複数分野語が関連の深い分野を大きく誤るようなことはない（すなわち、複数分野語の頻度の配分比率をある程度修正することができる）このようにして求められる共起ベクトルや文書共起ベクトルは、分類対象文書に現れた複数分野語の頻度値を、当該複数分野語の強関連分野に配分する際（図８のＳ６４、Ｓ６５）に用いられる。そして、この結果求められた文書ベクトルを用いて分類対象文書の分類先が決定される。この実施の形態２では、共起ベクトル及び文書共起ベクトルは、実施の形態１と比べて詳細化されているが、このほかの詳細化関連度テーブルや文書ベクトルは実施の形態１のものと同じでよい。

【０１３２】このように、実施の形態２によれば、共起ベクトルのスパースネス（基底の疎らさ）を吸収することが可能となり、実施の形態１で共起ベクトルでうまく処理できなかった分類対象文書に対しても、より好ましい分類結果を得ることができる。

【０１３３】以上、この発明の好適な実施の形態を説明した。以上に説明した各実施の形態は、あくまで一例に過ぎず、こほかにも様々なバリエーションが考えられる。

【０１３４】例えば、上記各実施の形態においては、文書から単語を切り出すために形態素解析を用いたが、この形態素解析の代わりに、平仮名、カタカナ、漢字、数字などの文字タイプの情報を利用して疑似的に単語分割を行うことも可能である。また、このほかにも言語処理分野で知られている様々な単語分割手法を適用することができる。

【０１３５】また、上記各実施の形態では単語と分野との関連度を、カイ２乗検定を応用した計算式により求めたが、関連度の求め方はこれに限らない。例えばTF・IDF(term frequency times inverse document frequency)など、統計分野で知られている様々な手法を利用することができる。

【０１３６】また、上記各実施の形態では、共起ベクトルを作成する際に、共起単語の調査範囲を同一段落とするほか、文書全体に範囲を広げたり、逆にその単語を含む文に範囲を限定したりすることが可能であると説明し

たが、このほかにも、注目する単語の前後所定文字数以内の範囲といった限定の仕方も可能である。これは、単語分割情報における単語位置の情報を用いて実現することができる。

【０１３７】また、分類対象文書と各分野との間の関連度を求めるための方法は、上記実施の形態１に示したベクトルの内積を利用する方法に限らず、様々な距離計算アルゴリズムを利用することが可能である。

【０１３８】また、上記実施の形態では、共起ベクトルを長さ１に正規化したが、次のような方法により、より共起ベクトルを特徴付けることもできる。すなわち、上記実施の形態１の方法により求められた各共起ベクトルにおいて、複数の共起ベクトルが、ある共通の単語に対して共に０でない成分値を有している場合は、その成分値をそれら共起ベクトルの数で除した値に置き換えるという方法である。例えば、図３１に示すように、実施の形態１の方法で（ａ）に示すような共起ベクトルが得られたとする。（ａ）の２つの共起ベクトルは、「総裁：名詞」に対応する成分値が共に０でない値となっているので、これを（ｂ）に示すようにそれぞれ２で割った値に置き換えるのである。ある単語が複数の共起ベクトルにおいて０でない成分値を持つということは、その単語はそれら複数の共起ベクトルに対応する分野同士の間での分類においては重要度が低いと考えることができるので、このような方法により共起ベクトルに特徴づけを行うことができる。

【０１３９】

【発明の効果】以上説明したように、この発明によれば、分類済み文書から複数分野語を学習し、この複数分野語に注目して関連度テーブルや分類対象文書の単語の頻度情報を詳細化するので、分類対象文書の各分野への関連度をより詳細に分析することができ、類似する分野間での分類の精度を向上させることができる。したがって、この発明によれば、細かい分類を行う場合でも分類精度の劣化を抑えることができる。

【０１４０】また、単語の強関連分野の判定基準となる閾値を、関連度テーブルにおける当該単語の各分野に対する関連度の中の最大値に基づき定めるので、各単語の関連度の分布に合わせ、相対的にみて関連度が高い分野を強関連分野として抽出することができる。

【０１４１】また、分類対象文書の出現単語の頻度情報の詳細化を、複数分野語の共起単語の傾向を示す共起ベクトルを利用して行うことにより、複数分野語が現れた周囲の状況に基づき当該複数分野語がどの分野の単語として現れたかを評価し、適切に詳細化を行うことができる。

【０１４２】また、各単語の概念的な階層関係を記述したシソーラスを用いて共起ベクトルや文書共起ベクトルを拡張することにより、共起情報のスパースネスを解消してより的確に単語が使われた状況を選択することができ、より的確な文書の自動分類を行うことができる。

【図面の簡単な説明】

【図１】　この発明の実施の形態１のシステムの構成図である。

【図２】　この発明の実施の形態１のシステムにおける全体的な処理手順を示すフローチャートである。

【図３】　単語分割／頻度抽出処理部の処理手順を示すフローチャートである。

【図４】　関連度演算部の処理手順を示すフローチャートである。

【図５】　複数分野語処理部による複数分野語検出処理の手順を示すフローチャートである。

【図６】　複数分野語処理部による詳細化関連度テーブルの作成処理の手順を示すフローチャートである。

【図７】　複数分野語処理部による共起ベクトル作成処理の手順を示すフローチャートである。

【図８】　分類対象文書の分類の手順を示すフローチャートである。

【図９】　分類済み文書の一例を示す図である。

【図１０】　文書分類先テーブルのデータ内容の一例を示す図である。

【図１１】　分類済み文書の頻度情報の一例を示す図である。

【図１２】　分類済み文書の単語分割情報の一例を示す図である。

【図１３】　頻度集計テーブルの一例を示す図である。

【図１４】　関連度テーブルの一例を示す図である。

【図１５】　複数分野語リストの一例を示す図である。

【図１６】　共起ベクトルの作成のために構築される共起頻度情報の一例を示す図である。

【図１７】　共起ベクトルの一例を示す図である。

【図１８】　詳細化頻度集計テーブルの一例を示す図である。

【図１９】　詳細化関連度テーブルの一例を示す図である。

【図２０】　分類対象文書の一例を示す図である。

【図２１】　分類対象文書の頻度情報の一例を示す図である。

【図２２】　分類対象文書の単語分割情報の一例を示す図である。

【図２３】　文書共起ベクトルの一例を示す図である。

【図２４】　分類対象文書の文書ベクトルに対応する各単語の頻度を示す表の一例を示す図である。

【図２５】　図２０の分類対象文書の各分野に対する関連度を、複数分野語を考慮した実施の形態１の手法により求めた結果を示す図である。

【図２６】　図２０の分類対象文書の各分野に対する関連度を、複数分野語を考慮しない手法で求めた結果を示す図である。

【図２７】　この発明の実施の形態２のシステムの構成

29

図である。

【図２８】　この発明の実施の形態２におけるシソーラス利用型複数分野語処理部の処理手順を示すフローチャートである。

【図２９】　実施の形態２におけるシソーラスの利用の仕方を説明するための図である。

【図３０】　シソーラスを用いた共起ベクトルの拡張処理を説明するための図である。

【図３１】　共起ベクトルの特徴付けの仕方の一例を示す図である。

【図３２】　従来の文書自動分類装置の構成図である。

【図３３】　図３２の従来装置における分類学習の結果の一例を示す図である。

【図３４】　別の従来の文書自動分類装置の構成図である。

【図３５】　図３４の従来装置で生成される単語ベクトルの一例を示す図である。

【図３６】　図３４の従来装置で生成される文書ベクトルの一例を示す図である。

30

【図３７】　更に別の従来の文書自動分類装置の構成図である。

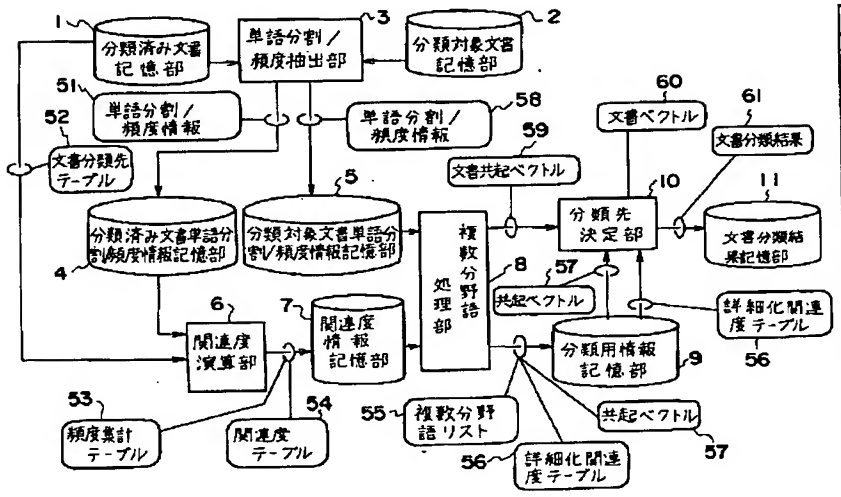
【図３８】　図３７の従来装置にて用いられる意味属性の情報の一例を示す図である。

【図３９】　図３７の従来装置において分類学習時に作成される各意味属性の頻度の集計結果を示す図である。

【符号の説明】

１　分類済み文書記憶部、２　分類対象文書記憶部、３　単語分割／頻度抽出部、４　分類済み文書単語分割／頻度情報記憶部、５　分類対象文書単語分割／頻度情報記憶部、６　関連度演算部、７　関連度情報記憶部、８　複数分野語処理部、９　分類用情報記憶部、１０　分類先決定部、１１　文書分類結果記憶部、２１　シソーラス利用型複数分野語処理部、２２　シソーラス、５１，５８単語分割／頻度情報、５２　文書分類先テーブル、５３　頻度集計テーブル、５４　関連度テーブル、５５　複数分野語リスト、５６　詳細化関連度テーブル、５７　共起ベクトル、５９　文書共起ベクトル、６０　文書ベクトル、６１　文書分類結果。

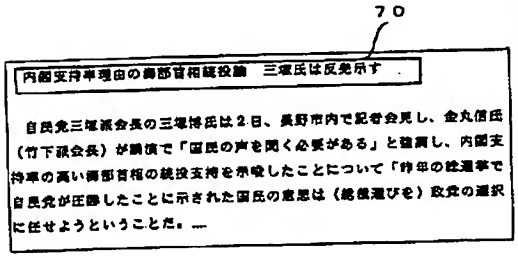
【図１】



【図１１】

単語	品詞	頻度
....		
会見	サ変名詞	1
会長	名詞	2
海部	名詞	2
記者	名詞	1
....		
国民	名詞	4
....		
首相	名詞	3
....		
総裁	名詞	3
....		
内閣	名詞	3
....		

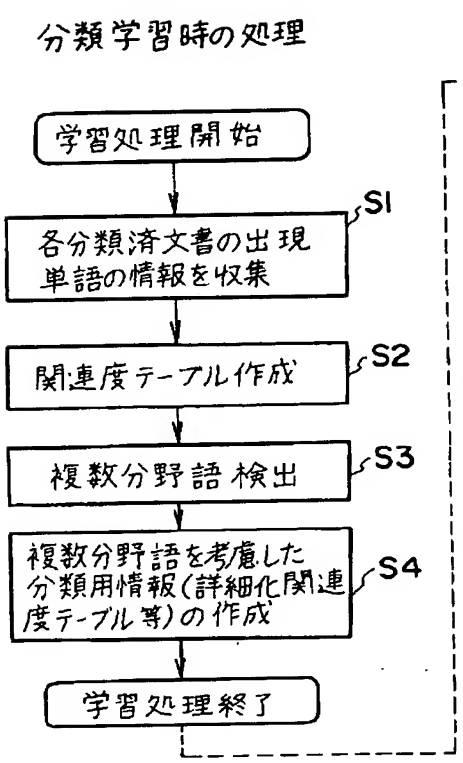
【図９】



【図１０】

文書名（ファイル名）	分野
...	
11/04M/04--09	政治
....	....
11/12M/09--06	経済
11/12M/09--08	政治、経済
....	....

【図２】



【図１５】

単語	品詞	強関連分野
....		
首相	名詞	政治、経済
....		
市場	名詞	経済、国際
....		

【図２１】

単語	品詞	頻度
....		
東京証券取引所	名詞	1
下半期	名詞	1
各社	名詞	2
株価	名詞	1
市場	名詞	4
....		
首相	名詞	3
....		
出来高	名詞	3
....		
薄商い	名詞	3
....		

【図２４】

単語	品詞	頻度
薄商い	名詞	3
市場(経済)	名詞	3.2
市場(国際)	名詞	0.8
....		
出来高	名詞	3
....		
首相(政治)	名詞	0.5
首相(経済)	名詞	2.5
....		

【図１２】

情報タイプ	単語位置	単語	品詞
P			
S			
W	3	内閣	名詞
W	7	支持率	名詞
W	13	理由	名詞
W	19	海部	名詞
W	23	首相	名詞
W	27	続投論	名詞
...			
P			
S			
W	55	自民党	名詞
W	61	三塚	名詞
W	67	会長	名詞
....			
W	99	記者	名詞
S			
W	544	首相	名詞
....			

【図１３】

単語	政治	経済	労働	...	スポーツ	国際
首相	50	30	3		1	10
...						
出来高	1	10	0		0	2
薄商い	0	3	0		0	0
市場	0	10	0		0	8
...						

【図１４】

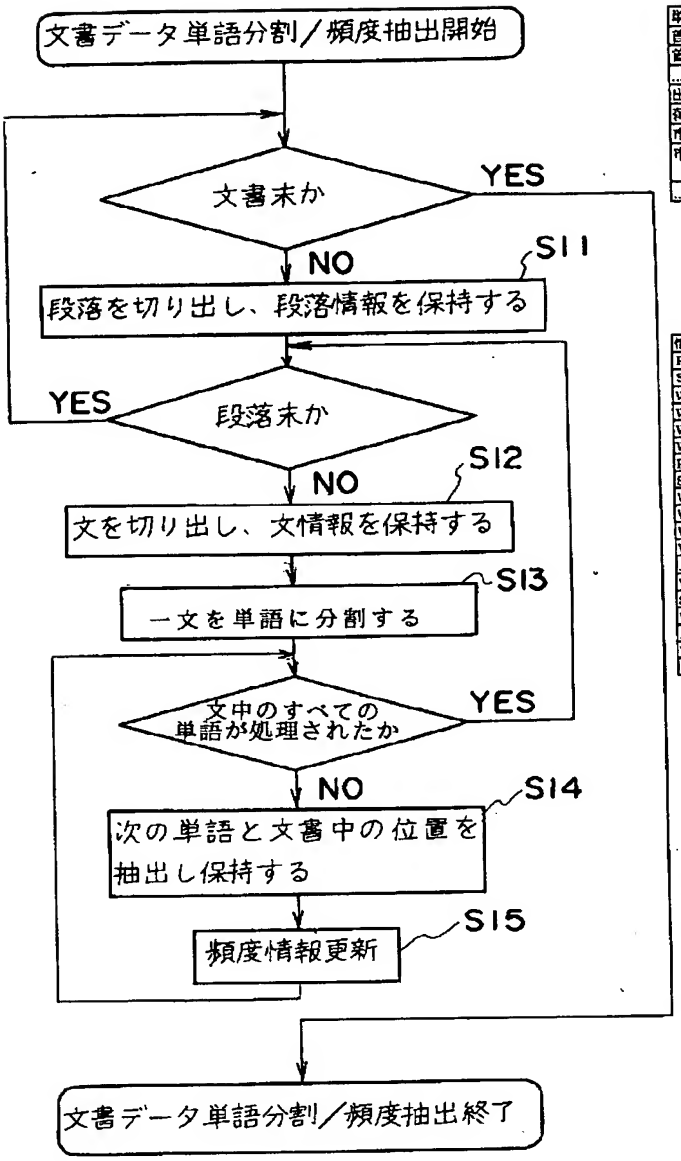
単語	政治	経済	労働	...	スポーツ	国際
首相	66.7	20.8	0.0		-1.6	1.6
...						
出来高	-3.2	9.0	-1.0		-1.0	0.0
薄商い	-1.6	10.0	-0.3		-0.3	-0.6
市場	-5.0	9.0	-1.0		-1.0	18.0
...						

【図２３】

首相　＝（2,　0,　0,　4,　1,　...）

内閣　国民　総裁　市場　経済
(名詞) (名詞) (名詞) (名詞) (名詞)

【図３】



【図１８】

単語	政治	経済	労働	…	スポーツ	国際
首相(政治)	30	0	1.5		0.5	5
首相(経済)	0	30	1.5		0.5	5
…						
出来高	1	10	0		0	2
薄商い	0	5	0		0	0
市場(経済)	0	10	0		0	0
市場(国際)	0	0	0		0	8
…						

【図２２】

情報タイプ	単語位置	単語	品詞
P			
S			
W	11	薄商い	名詞
W	27	出来高	名詞
W	51	東証	名詞
W	55	1部	名詞
P			
S			
W	63	東京証券取引所	名詞
W	77	市場	名詞
W	81	1部	名詞
W	95	出来高	名詞
…			
W	192	水準	名詞
S			
W	207	市場	名詞
…			
W	223	首相	名詞
…			

【図２５】

計算結果

政治：　160.0×0.5+(-6.6)×2.5+(-3.2)×3+(-1.6)×3+(-2.6)×3.2+(-2.3)×0.8+…　=38.9+…

経済：　(-6.0)×0.5+114.0×2.5+9.0×3+10.6×3+29.0×3.2+(-1.6)×0.8+…　=429.3+…

労働：　(-0.1)×0.5+(0.0)×2.5+(-1.0)×3+(-0.3)×3+(-0.5)×3.2+(-0.4)×0.8+…　=-5.8+…

スポーツ：　(-1.1)×0.5+(-0.5)×2.5+(-1.0)×3+(-0.3)×3+(-0.5)×3.2+(-0.4)×0.8+…　=-7.6+…

国際：　0.2×0.5+2.0×2.5+0.0×3+0.8×3+(-1.0)×3.2+53.5×0.8+…　=42.9+…

【図１９】

単語	政治	経済	労働	…	スポーツ	国際
首相(政治)	160.0	-8.0	-0.1		-1.1	0.2
首相(経済)	-6.6	114.0	0.0		-0.5	2.0
…						
出来高	-3.2	9.0	-1.0		-1.0	0.0
薄商い	-1.6	10.0	-0.3		-0.3	0.6
市場(経済)	-2.6	29.0	-0.5		-0.5	-1.0
市場(国際)	-2.3	-1.8	-0.4		-0.4	53.5
…						

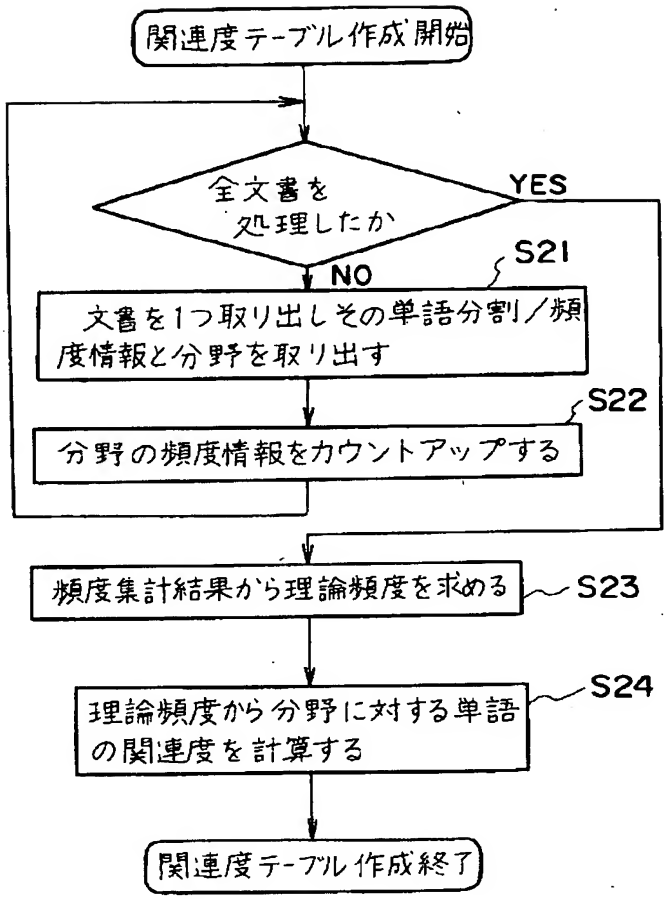
【図２０】

記録的な薄商い　１１日の出来高１億５０００万株　東証１部

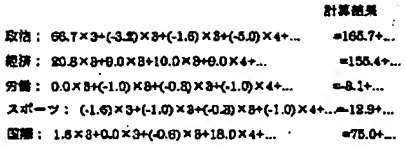
東京証券取引所市場１部の１１日の出来高は、１億５４３７万株と１日の立ち会いでは１９８４年５月２８日（１億３５６３万株）以来の低い水準となった。市場関係者は官択首相就任と同時期に公定歩合引き下げがあると期待していたが、早くても今週後半になったた　、市場の見送り気分が強まった。９月中間決算で赤字や大幅減益となった証券各社は、年度下半期の東証１部の出来高を１日当たり４、５億株と見込んで予算を立てており、薄商いが続くと、証券各社の苦境は深刻さを増しそうだ。

同日の１銘の平均株価は先週末終値に比べて２５３円５０銭安の２万４

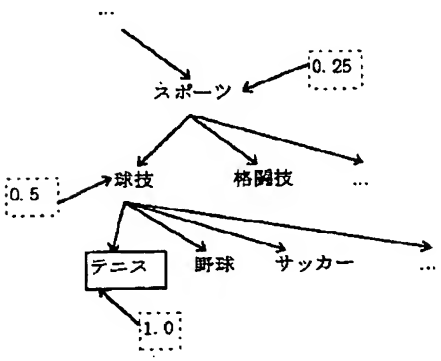
【図４】



【図２６】



【図２９】



【図３３】

見出し文字列	品詞	題名	著者	目的	...	合計	分類名
自然語	名詞	0	1	1		17	自語処理
形態	名詞	1	2	1		21	自語処理
LR解析	名詞	0	0	0		4	自語処理
....							
AD変換	名詞	1	1	1		11	電気回路
アナログ	名詞	0	2	1		12	電気回路
....							

【図３５】

(アメリカ, 政府, 先進, 主要, 国, ココム, ..., 意向

アメリカ＝(1, 1, 1, 1, 1, 1, ..., 0)

兵器 ＝(0, 0, 0, 0, 1, 1, ..., 1)

【図１７】

（ａ）

首相(政治)＝（5，　3，　5，　0，　0，　...）

首相(経済)＝（0，　0，　3，　5，　8，　...）

内閣　国民　総裁　市場　経済
(名詞) (名詞) (名詞) (名詞) (名詞)

↓正規化

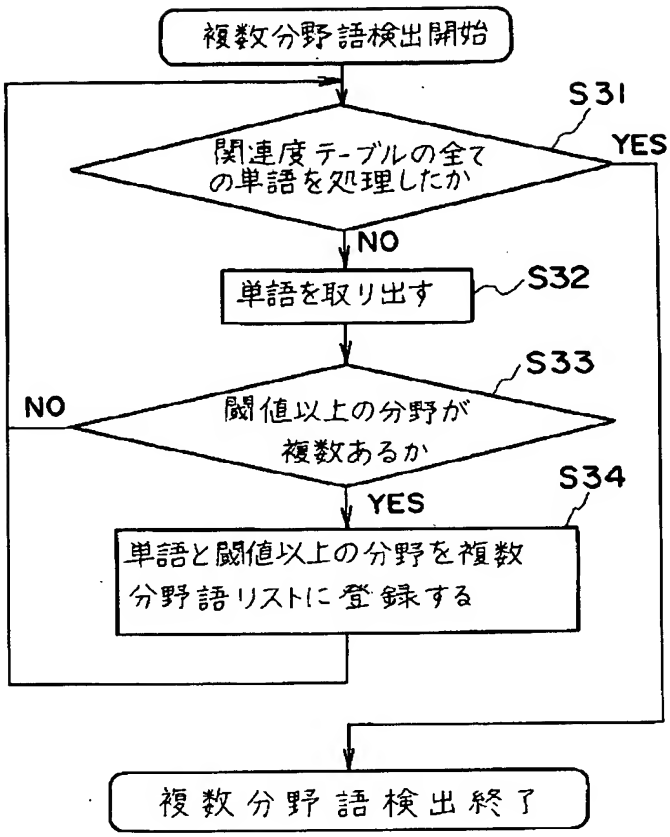
（ｂ）

首相(政治)＝（0.41，0.25，0.41，　0，　0，　...）

首相(経済)＝（0，　0，　0.33，　0.55，　0.77，　...）

内閣　国民　総裁　市場　経済
(名詞) (名詞) (名詞) (名詞) (名詞)

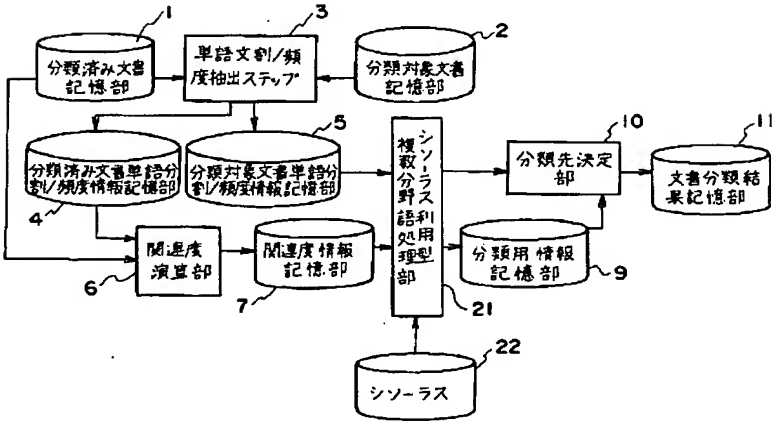
【図５】



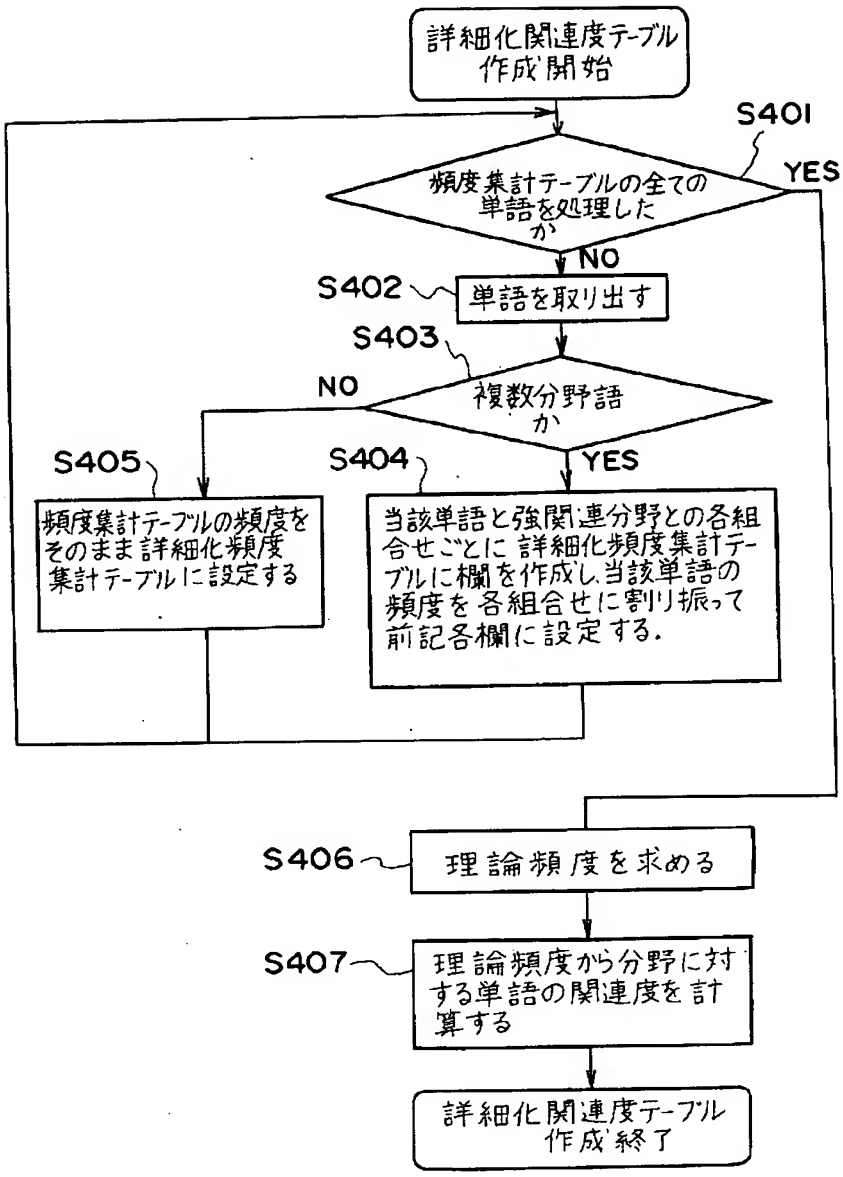
【図３６】

	(アメリカ, 政府, 先進, 主要, 国, ココム, ..., 意向
アメリカ＝(1,	1, 1, 1, 1, 1, ..., 0)
兵器　＝(0,	0, 0, 0, 1, 1, ..., 1)
文書ベクトル＝(1,	1, 1, 1, 2, 2, ..., 1)

【図２７】



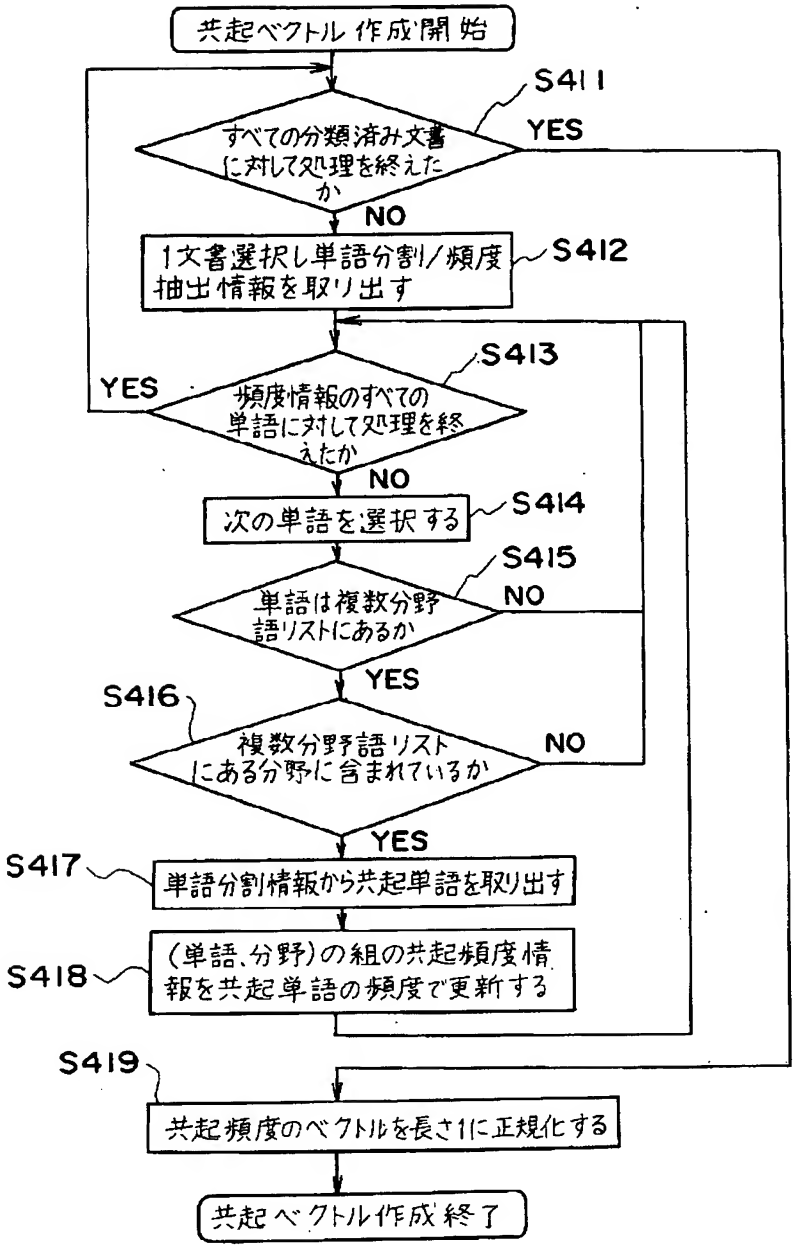
【図６】



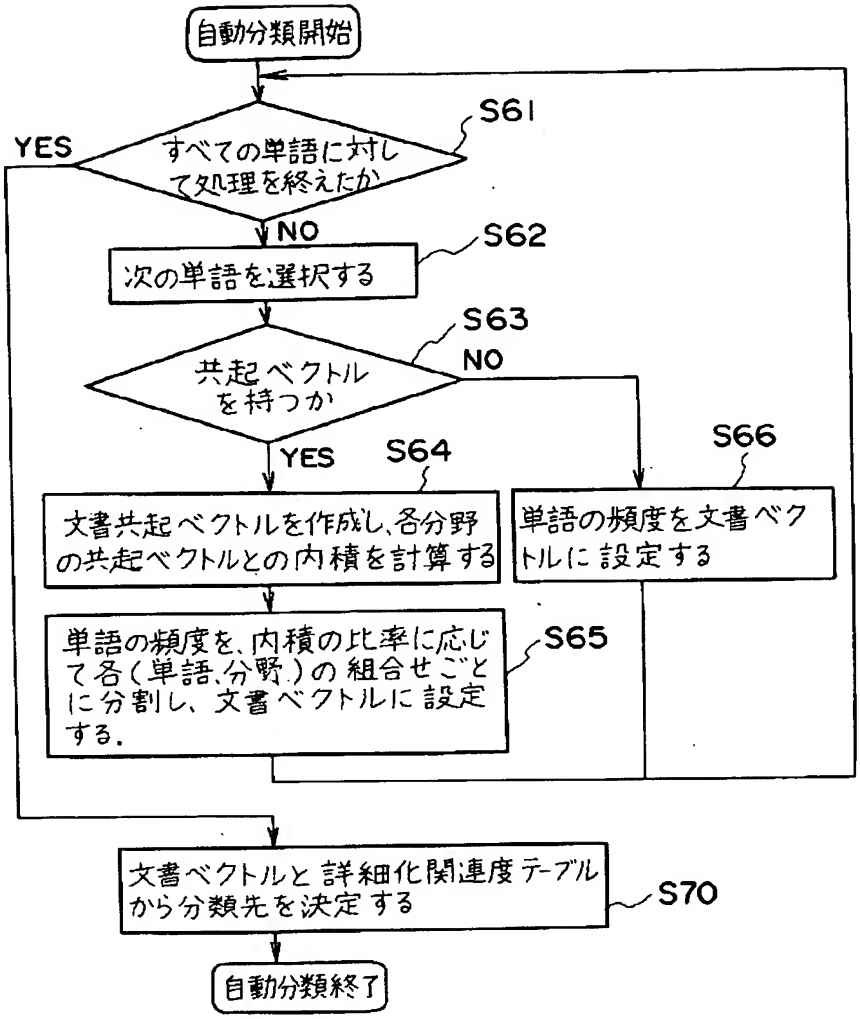
【図３８】

名詞	一般名詞意味属性	固有名詞意味属性
醤油（一般名詞）	[調味料]	−
おかず（一般名詞）	[食料]	−
秋田（固有名詞）	[行政区画]、[学校]	[県名]、[市名]、.....

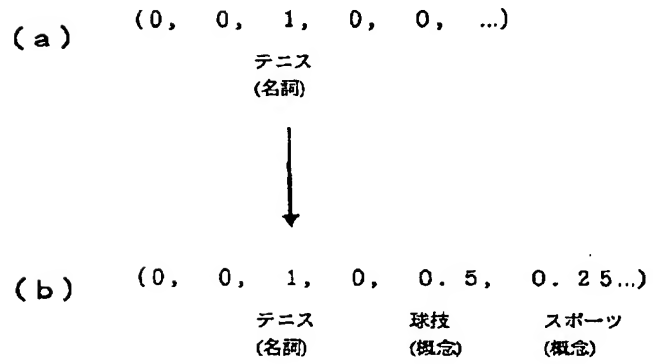
【図７】


共起ベクトル作成開始
S411
すべての分類済み文書に対して処理を終えたか
YES
NO
1文書選択し単語分割/頻度抽出情報を取り出す
S412
S413
YES
頻度情報のすべての単語に対して処理を終えたか
NO
次の単語を選択する
S414
S415
単語は複数分野語リストにあるか
NO
YES
S416
複数分野語リストにある分野に含まれているか
NO
YES
S417
単語分割情報から共起単語を取り出す
S418
(単語、分野)の組の共起頻度情報を共起単語の頻度で更新する
S419
共起頻度のベクトルを長さ1に正規化する
共起ベクトル作成終了

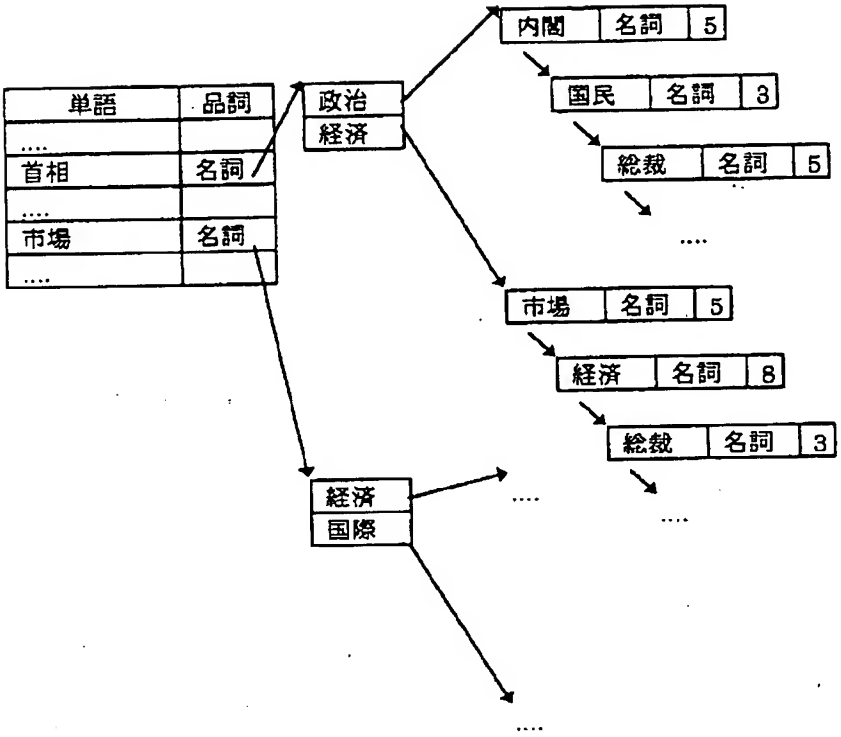
【図８】


自動分類開始
S61
すべての単語に対して処理を終えたか
YES
NO
S62
次の単語を選択する
S63
共起ベクトルを持つか
NO
YES
S64
S66
文書共起ベクトルを作成し、各分野の共起ベクトルとの内積を計算する
単語の頻度を文書ベクトルに設定する
単語の頻度を、内積の比率に応じて各（単語、分野）の組合せごとに分割し、文書ベクトルに設定する。
S65
文書ベクトルと詳細化関連度テーブルから分類先を決定する
S70
自動分類終了

【図３０】


(a)
(0, 0, 1, 0, 0, ...)
テニス
(名詞)
(b)
(0, 0, 1, 0, 0.5, 0.25...)
テニス
(名詞)
球技
(概念)
スポーツ
(概念)

【図１６】



【図３１】

（ａ）

首相(政治)＝（0.41，0.25，0.41，　0，　0，　...）

首相(経済)＝（0，　0，　0.33，　0.55，　0.77，　...

内閣 (名詞)　国民 (名詞)　総裁 (名詞)　市場 (名詞)　経済 (名詞)

（ｂ）

首相(政治)＝（0.41，0.25，0.20，　0，　0，　...）

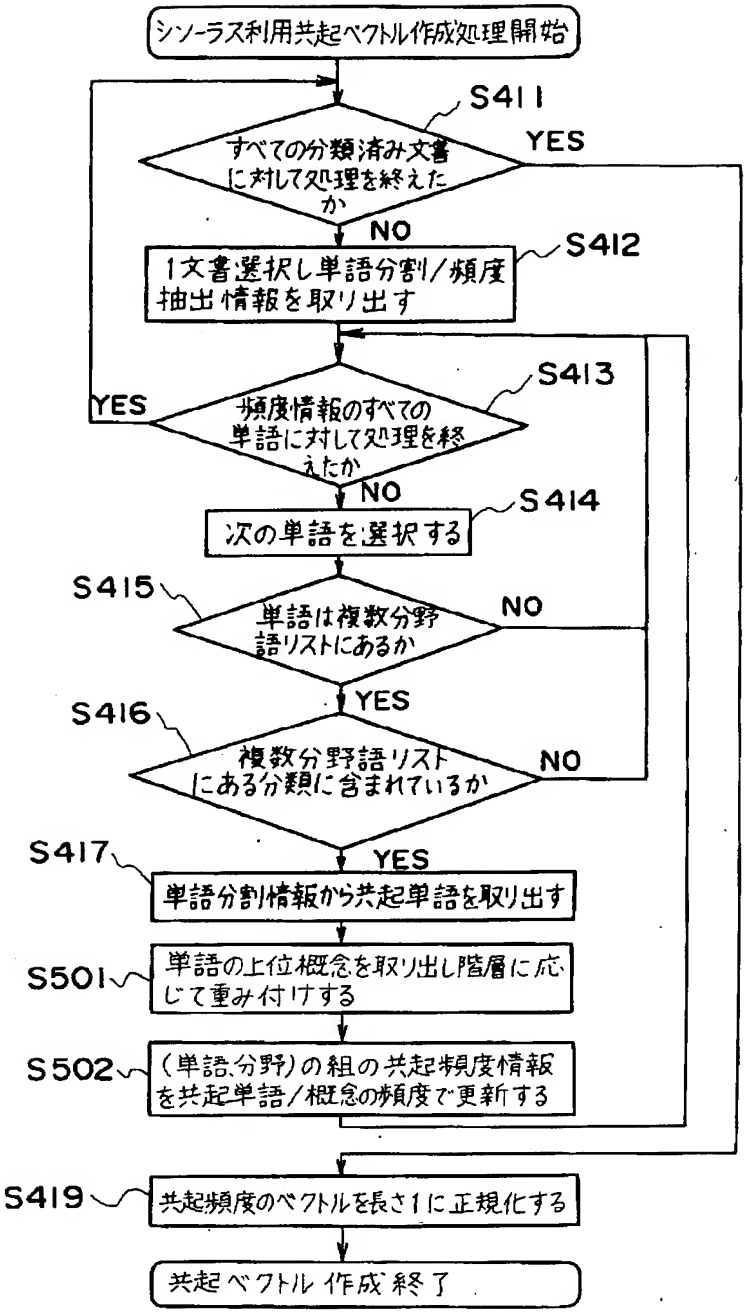
首相(経済)＝（0，　0，　0.16，　0.55，　0.77，　...

総裁(名詞)

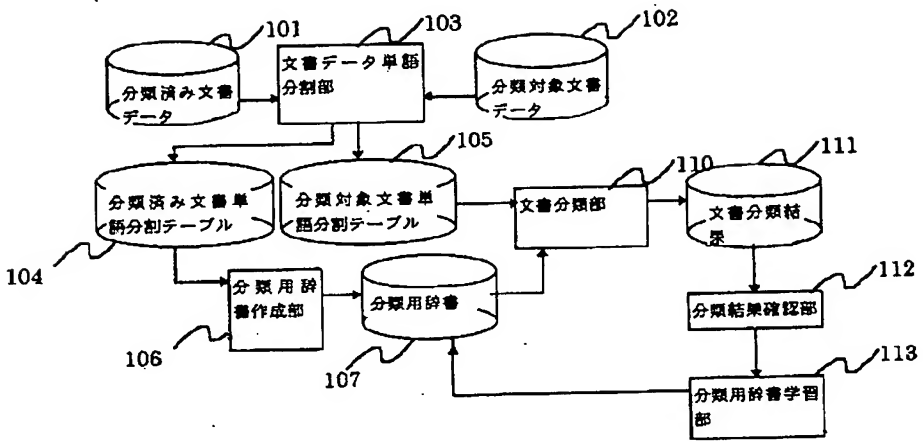
【図３９】

	分野				
意味属性	政治	経済	....	家庭	運輸通信
[人工物]	7	10		12	12
[資材]	0	7		1	9
.....					
[食料]	1	1		9	1
[調味料]	0	0		7	1

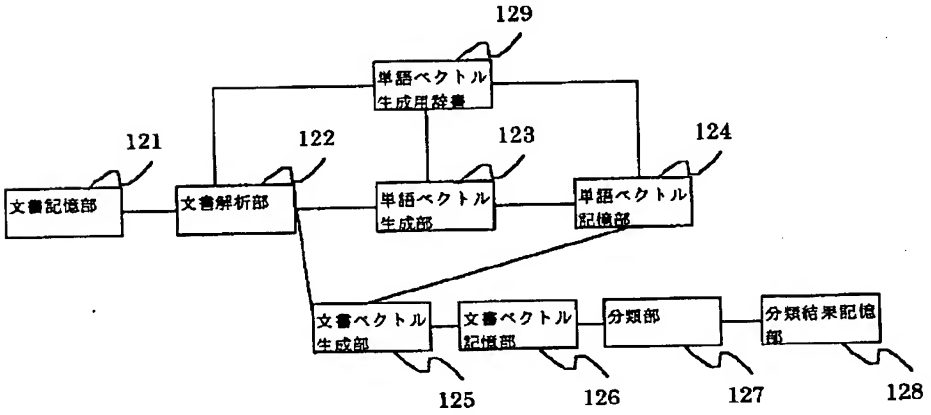
【図２８】


シソーラス利用共起ベクトル作成処理開始
S411
すべての分類済み文書に対して処理を終えたか
YES
NO
S412
1文書選択し単語分割/頻度抽出情報を取り出す
S413
頻度情報のすべての単語に対して処理を終えたか
YES
NO
S414
次の単語を選択する
S415
単語は複数分野語リストにあるか
NO
YES
S416
複数分野語リストにある分類に含まれているか
NO
YES
S417
単語分割情報から共起単語を取り出す
S501
単語の上位概念を取り出し階層に応じて重み付けする
S502
(単語,分野)の組の共起頻度情報を共起単語/概念の頻度で更新する
S419
共起頻度のベクトルを長さ1に正規化する
共起ベクトル作成終了

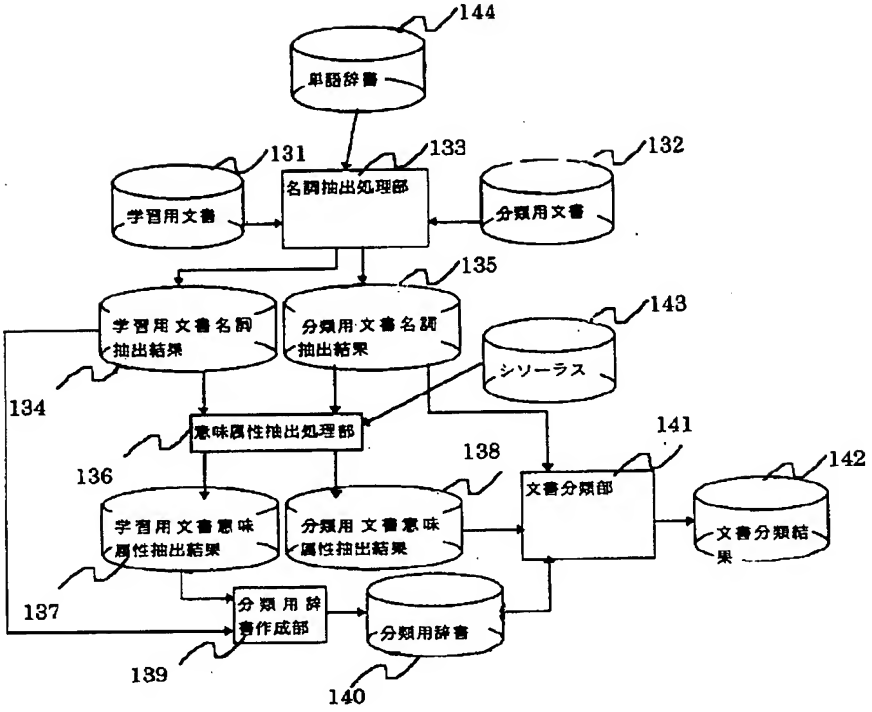
【図３２】


101
103
102
分類済み文書データ
文書データ単語分割部
分類対象文書データ
105
111
110
分類済み文書単語分割テーブル
分類対象文書単語分割テーブル
文書分類部
文書分類結果
104
分類用辞書作成部
分類用辞書
112
分類結果確認部
106
107
113
分類用辞書学習部

【図３４】


129
単語ベクトル生成用辞書
121
122
123
124
文書記憶部
文書解析部
単語ベクトル生成部
単語ベクトル記憶部
文書ベクトル生成部
文書ベクトル記憶部
分類部
分類結果記憶部
125
126
127
128

【図３７】



フロントページの続き

(72)発明者　高山　泰博
東京都千代田区丸の内二丁目２番３号　三
菱電機株式会社内